

RESEARCH OUTPUTS / RÉSULTATS DE RECHERCHE

Industrial and Medical Anomaly Detection Through Cycle-Consistent Adversarial Networks

Bougaham, Arnaud; Delchevalerie, Valentin; Adoui, Mohammed El; Frénay, Benoît

Publication date:
2023

[Link to publication](#)

Citation for published version (HARVARD):

Bougaham, A, Delchevalerie, V, Adoui, ME & Frénay, B 2023 'Industrial and Medical Anomaly Detection Through Cycle-Consistent Adversarial Networks'.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Industrial and Medical Anomaly Detection Through Cycle-Consistent Adversarial Networks

Arnaud Bougaham*

Fac. of Computer Science

NaDI institute

University of Namur

Namur, Belgium

arnaud.bougaham@unamur.be

Valentin Delchevalerie*

Fac. of Computer Science

NaDI & naXys institutes

University of Namur

Namur, Belgium

valentin.delchevalerie@unamur.be

Mohammed El Adoui

Fac. of Computer Science

NaDI institute

University of Namur

Namur, Belgium

mohammed.eladoui@unamur.be

Benoît Frénay

Fac. of Computer Science

NaDI institute

University of Namur

Namur, Belgium

benoit.frenay@unamur.be

Abstract—In this study, a new Anomaly Detection (AD) approach for real-world images is proposed. This method leverages the theoretical strengths of unsupervised learning and the data availability of both normal and abnormal classes. The AD is often formulated as an unsupervised task motivated by the frequent imbalanced nature of the datasets, as well as the challenge of capturing the entirety of the abnormal class. Such methods only rely on normal images during training, which are devoted to be reconstructed through an autoencoder architecture for instance. However, the information contained in the abnormal data is also valuable for this reconstruction. Indeed, the model would be able to identify its weaknesses by better learning how to transform an abnormal (or normal) image into a normal (or abnormal) image. Each of these tasks could help the entire model to learn with higher precision than a single normal to normal reconstruction. To address this challenge, the proposed method utilizes Cycle-Generative Adversarial Networks (Cycle-GANs) for abnormal-to-normal translation. To the best of our knowledge, this is the first time that Cycle-GANs have been studied for this purpose. After an input image has been reconstructed by the normal generator, an anomaly score describes the differences between the input and reconstructed images. Based on a threshold set with a business quality constraint, the input image is then flagged as normal or not. The proposed method is evaluated on industrial and medical images, including cases with balanced datasets and others with as few as 30 abnormal images. The results demonstrate accurate performance and good generalization for all kinds of anomalies, specifically for texture-shaped images where the method reaches an average accuracy of 97.2% (85.4% with an additional zero false negative constraint).

Index Terms—Cycle-GANs, Industry 4.0, Industrial Images, Medical Images, Anomaly Detection, Zero False Negative

I. INTRODUCTION

This work proposes a new approach with a Generative Adversarial Networks (GAN) architecture for the task of Anomaly Detection (AD), which aims to combine the advantages of both unsupervised learning and the data availability of the normal and abnormal classes. Indeed, AD is often formulated as an unsupervised task due to the frequent high imbalance between normal and abnormal data, and the need for generalization across a wide range of anomalies. Therefore, in an autoencoder architecture for instance, the AD method is trained by reconstructing normal data only. Nevertheless,

valuable information is missing during the training step. Only the normal class is taken into account, and the reconstruction from the abnormal to the normal class is not included in this learning process. Yet, this is precisely the task we are expecting from an AD method during the inference step. The proposed method seeks to overcome this limitation by learning how to transform an abnormal image into a normal image by exploiting samples from both classes. The objective is to generate a reconstructed image where any abnormal pixel is replaced by a normal one in a visually-coherent manner. During the training step, the “normal generator” is constrained by another “abnormal generator” in an adversarial framework, using Cycle-Generative Adversarial Networks (Cycle-GANs) [1]. Also, reconstructing the abnormal data during the learning step yields to a better normal generator than the classical methods using only the normal class. Even if the abnormal datasets can be small (as it is usually the case in the AD context), the normal generator performs better, because its performance is also constrained by the abnormal generator, resulting in a good reconstruction. We still consider this as an unsupervised learning task because the abnormal data used during training is not necessarily representative of all anomalies that could occur. Abnormal data are just given to help during the training phase by giving more feedback to the generators. Therefore, the generalization is guaranteed as it is the case in a classical GAN context, except that the normal reconstruction is much less noisy.

Cycle-GAN is a well-known architecture proposed a few years ago. It constitutes an elegant way to learn conditional mappings from two different domains \mathcal{X} and \mathcal{Y} (for image-to-image translation) by applying a cycle-consistent constraint on the transformations. The popularity of cycle-GANs lies in the fact that they only need a dataset of unpaired images to learn the mappings. In other words, they do not need the one-to-one correspondence between data from \mathcal{X} and \mathcal{Y} , but only two independent sets of data $\{x_i \in \mathcal{X}\}$ and $\{y_i \in \mathcal{Y}\}$. As an example, one can consider two unpaired datasets $\{x_i\}$ and $\{y_i\}$ made of unrelated aerial images and Google maps, respectively. A cycle-GAN can be trained to learn meaningful mappings from \mathcal{X} to \mathcal{Y} and \mathcal{Y} to \mathcal{X} . Fig. 1 presents an example generated with this cycle-GAN.

* The first two authors contributed equally.



Fig. 1: Example generated from a cycle-GAN (see Section V for training details) that learns mappings between aerial photos \mathcal{X} and Google maps \mathcal{Y} (dataset from [2]). The initial image $x \in \mathcal{X}$ can be mapped to $\tilde{y} \in \mathcal{Y}$ thanks to a first generator G . The second one F can then go back from $\tilde{y} \in \mathcal{Y}$ to $\tilde{x} \in \mathcal{X}$. A cycle-consistent constraint enforces \tilde{x} to be close to x .

Many works have shown that cycle-GANs can be used in diverse image analysis tasks. Nonetheless, cycle-GANs remain seldom used in practice to solve problems in the industrial and medical areas. It is for example the case of AD where, to the best of our knowledge, no prior work directly exploits cycle-GANs. Compared to the state-of-the-art for AD in images, cycle-GANs seem to be more suitable than concurrent methods that rely on simple GANs. Furthermore, we show that the formalism of a cycle-GANs makes them really efficient and well-suited for AD. Moreover, the use of an identity loss allows the reconstruction of the normal-to-normal (and abnormal-to-abnormal) generator to be much less noisy than traditional GANs, making it possible to better discriminate normal and abnormal images. To illustrate this, we focus our experiments on several industrial and medical problems of AD. This is motivated by (i) the abundance of AD problems in these domains, and (ii) the positive societal impact of developing efficient AD algorithms for them. From our results, it clearly appears that cycle-GANs can be very efficient for specific types of images.

The main contributions of our work are as follow:

- Utilize abnormal data in the learning process to reinforce the normal generator, by reconstructing from both classes.
- Consider a cycle-GAN architecture, broadly used for image generation, as a powerful approach for AD.
- Apply an identity loss that allows a better discrimination between normal and abnormal images.
- Characterize and discuss the performances of the method for diverse industrial and medical AD problems.
- Give insights on the reasons why cycle-GANs fit well to the problem of AD for specific natures of images, and consider further investigations on the use of cycle-GANs in the industrial and medical domains.

First, we present the theoretical prerequisites to understand cycle-GANs in Section II. After that, Section III presents the previous works, and highlights that most of them only use simple GAN architectures by training with only normal data. The proposed AD method with cycle-GANs is described in Section IV. Section V then introduces the considered datasets, the experimental setup as well as the results. A discussion and

some limitations with the use of cycle-GANs are presented in Section VI before concluding with Section VII.

II. BACKGROUND ON CYCLE-GANs

This section introduces the formalism for cycle-GANs. The building blocks and the loss functions are described.

A. Building Blocks

Cycle-Generative Adversarial Networks (Cycle-GANs) learn image-to-image mappings from an unpaired dataset constituted of two types of images from domains \mathcal{X} and \mathcal{Y} . Cycle-GANs are obtained by tying together two distinct conditional GANs with a cycle-consistent constraint. The first GAN is made of a generator

$$G : \mathcal{X} \cup \mathcal{Y} \longrightarrow \mathcal{Y} : G(z) = \tilde{y}, \quad (1)$$

and a discriminator D_Y , and the other is made of a generator

$$F : \mathcal{X} \cup \mathcal{Y} \longrightarrow \mathcal{X} : F(z) = \tilde{x}, \quad (2)$$

and a discriminator D_X . For convenience, let's already consider an AD task where \mathcal{X} are abnormal images, while \mathcal{Y} are normal ones. On the one hand, the aim of G is to generate from $x \in \mathcal{X} \cup \mathcal{Y}$ an image such that D_Y cannot distinguish it from real normal images in \mathcal{Y} . On the other hand, F aims to generate images such that D_X is fooled and cannot distinguish it from real abnormal images in \mathcal{X} . To achieve this, cycle-GANs are trained by optimizing a combination of different losses that are described in the next section.

B. Objective Function

A cycle-GAN requires to train two GANs, and to tie them together with a cycle-consistent constraint. The loss can be broken down into three parts such that

$$G^*, F^* = \arg \min_{F, G} \max_{D_X, D_Y} \mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}} \mathcal{L}_{\text{cyc}} + \lambda_{\text{ide}} \mathcal{L}_{\text{ide}}, \quad (3)$$

where λ_{cyc} and λ_{ide} are meta-parameters that constraint the different parts of the loss.

The first part is made of two classical adversarial losses [3]

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{GAN}}(G, D_Y) + \mathcal{L}_{\text{GAN}}(F, D_X), \quad (4)$$

where,

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_y [\log(D(y))] + \mathbb{E}_x [\log(1 - D(G(x)))]. \quad (5)$$

On the one side, by enforcing G (resp. F) to minimize \mathcal{L}_{adv} , the generator will try to generate images that look similar to images from \mathcal{Y} (resp. \mathcal{X}). On the other side, by enforcing D_Y (resp. D_X) to maximize \mathcal{L}_{adv} , the discriminator will try to distinguish between images coming from the generator G (resp. F) and real images in \mathcal{Y} (resp. \mathcal{X}).

The second part is motivated by the fact that the reconstructed images $F(G(x))$ and $G(F(y))$ should be close to x and y , respectively (i.e., the pair of GANs should be cycle-consistent). This is achieved by the cycle-consistent loss

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_x [\|F(G(x)) - x\|_1] + \mathbb{E}_y [\|G(F(y)) - y\|_1], \quad (6)$$

where L1 norm is used like in the original work on cycle-GANs [1].

In addition to the \mathcal{L}_{adv} and \mathcal{L}_{cyc} , an identity loss is added to constrain the generators to leave the images unmodified if they are already in the desired output domain, defined as

$$\mathcal{L}_{ide} = \mathbb{E}_x [\|F(x) - x\|_1] + \mathbb{E}_y [\|G(y) - y\|_1], \quad (7)$$

so as to enforce $F(x) = x$ and $G(y) = y$. In other words, F should not add anomalies if the input image is already abnormal, and G should not make any modification if it is already normal. Although the identity loss is present in the implementation of cycle-GANs from the original paper, it is not discussed and seldom used in practice. However, in the context of AD, the use of the identity loss is particularly relevant. Indeed, it is expected from G that it erases any abnormal pixel from the image. Nonetheless, in the case the image does not contain any of them, it should learn to leave it unmodified. This important property is exactly the one enforced by the identity loss. At the inference time, it also allows us to only use one of the two generators (the one that goes from abnormal to normal data, i.e., G) as potential abnormal pixels are revealed by comparing the reconstructed image with the original one.

III. RELATED WORKS

AD has long been an area of great concern in a wide range of fields such as biomedical [4], industrial [5] and security [6], [7]. Furthermore, a significant number of works have been published to characterize the AD approaches in the literature. The scope of this section is focused on previous works based on GANs and cycle-GANs, applied to the industrial and medical domains. GANs are used for many image-related tasks such as AD [8]–[11], segmentation, data augmentation, etc. However, to the best of our knowledge, cycle-GANs have been mostly used for data augmentation [12]–[14]. Since the purpose of our research is focused on the AD in medical and industrial applications, in this section, we review the most relevant researches applied to these two fields. From an in-depth analysis of the state-of-the-art methods associated with our research issue, the recent AD methods are mainly using only GANs. Yet, cycle-GANs can be highly useful for AD thanks to the combination of unsupervised learning and the data availability of the normal and abnormal classes.

Regarding the industrial studies, Bougaham et al. [5] propose to use intermediate patches for the inference step after a Wasserstein GAN learning process. The objective is to produce an efficient approach for AD on real industrial images of electronic Printed Circuit Board Assembly (PCBA). The technique can be used to assist current industrial image processing algorithms and to avoid tedious manual processing. Nevertheless, due to the wide variety of possible anomalies in a PCBA and the high complexity of autoencoder architecture, a real-world implementation remains a challenging task, specifically for small anomalies even if the method evolved to overcome some limitations in [11]. Rippel et al. [15] and Zhang et al. [16] suggest to use cycle-GANs to perform data augmentation by

generating synthetic images for industrial inspection. Recently, J. Liu et al. [17] develop an autoencoder technique of AD using images of aluminum surfaces. The challenge of this work is to detect the manufacturing errors using unlabeled data. They introduced a dual prototype loss approach to encourage encoder generated feature vectors to match their own prototype. Therefore, the root mean square error between feature vectors is used as an indicator of anomalies.

Regarding the medical field, Schlegl et al. [8] proposed an unsupervised AD framework GANs (f-AnoGAN), that can detect the unseen anomalies of medical subjects after being trained on healthy tomography images. On the other side, among the previous studies in medical imaging, authors of [18], [19] could be cited. These authors use cycle-GANs to perform the data augmentation task for MRI and CT scan images, respectively. Again, they show that using cycle-GANs for data augmentation leads to better segmentation performances afterwards.

Despite being better than the previous approaches, these deep AD approach use unsupervised deep learning techniques, such as autoencoders and GANs, to characterize the normal class, without using the insights given by the anomalies. In contrast, in this work, the anomaly images are leveraged immediately in the training phase, using it as prior knowledge to strengthen the model at recognizing anomalies. Furthermore, we check our method in both industrial and medical images while enforcing zero false negative (ZFN).

In a concise way, a thorough analysis of the most important studies in the literature shows that in the industrial and medical domains, cycle-GANs have been mainly used for data augmentation. This paper demonstrates the suitability of cycle-GANs for AD, in particular, for industrial and medical images, which, to our knowledge, has never been covered in the literature.

IV. METHODS

This section introduces the developed approach and shows its relevance for AD with cycle-GANs. The architecture for the training and inference steps is illustrated in Fig. 2.

The basic idea behind the use of cycle-GANs for AD is to exploit the conditional mapping learned by one of the two generators: the one that goes from abnormal to normal images. Indeed, by forward-propagating an abnormal image in this generator, it is expected to obtain a new image where the anomaly is erased. Nonetheless, and thanks to the identity loss, if a normal image is forward-propagated in the generator, it is expected to remain unchanged. Therefore, by comparing the output of the generator with its input, anomalies in the input images can be located. The other generator (normal-to-abnormal) is not really useful from a practical point of view. It is only useful to jointly train the first one, similarly to the two discriminators, but not for the AD inference step.

To perform the AD task, the normal and abnormal test images are given to the learned abnormal-to-normal generator. Then, an anomaly score is computed to measure the distance between the original test image and the reconstructed one. In

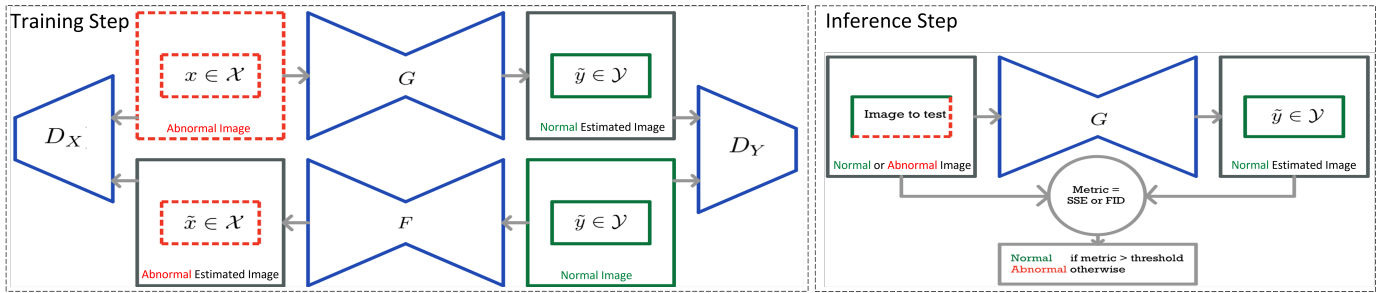


Fig. 2: Figure inspired from [1] presenting the architecture of the training (left side) and the inference (right side) steps. During the training step, the first generator G tries to map abnormal to normal images by fooling the discriminator D_Y that should not detect fake images. F and D_X follow the same idea but for normal images as input. During the inference step, only G is used even if the input can either be normal or abnormal images. This is possible thanks to the identity loss that enforces G to only modify abnormal pixels if any, and leaves the image unmodified otherwise.

this paper, two metrics are considered: a per-pixel sum of the squared differences (SSE), and a Fréchet Inception Distance (FID) [20]. The FID anomaly score is more elaborated and focuses on perceptual differences thanks to the use of a pre-trained Inception V3 network [21]. For each anomaly score, its potential for AD with cycle-GANs is assessed by building an anomaly detector with two different thresholds. The first threshold is set by minimizing the number of classification errors, which yields an anomaly detector with maximum accuracy (ACC). The second one is set so that all true positives are detected, i.e., only false alarms can be raised but no anomaly can be missed. This setting yields an anomaly detector with zero false negative (ZFN), which is the most useful in the business applications of the industrial or medical fields, where false negatives have large consequences for customers or patients. In summary, four anomaly detectors are built, i.e., one for each pair of metrics and thresholds. Note that the thresholds are set on the test sets, and make it possible to assess how much the two distributions (SSE and FID for normal and abnormal data) are discriminated. The use of an additional validation set should be preferred, but it would have been too costly in terms of abnormal data for several datasets (due to the scarcity of abnormal data available). Therefore, the accuracy values are an overestimate of the classification performances, and should be seen as a metric to quantify how much the method highlights abnormal images compare to normal images on the test sets. Our goal is indeed to measure the discriminative power of cycle-GANs for AD, so as to prospectively validate the practical interest of our idea in the industrial and medical domains.

V. EXPERIMENTS

This section presents the experiments carried out to evaluate the proposed AD method. First, the datasets and the data pre-processing steps are introduced. Next, the model architecture used for all the experiments is detailed. Finally, qualitative and quantitative results are presented, then discussed in Section VI.

A. Datasets

Eight datasets are used for both the industrial and medical domains, where several types of anomalies may occur. To assess the strengths and weaknesses of our method, we defined four categories of anomalies: small/large object-shaped, or small/large textured-shaped anomalies. Indeed, the anomaly can either arise on a specific object characterized by abrupt changes in pixels intensity (a screw for instance), or where the pixels intensity changes are more progressive and more homogeneous (in the structure of wood for instance).

To cover the industrial side, the public MVTEC-AD dataset [22] is used. It consists of different high resolution industrial images from 15 different categories of object and texture-shaped products with and without anomalies. In this work, 4 datasets were selected from MVTEC-AD to cover the different natures of images: the Screw (small object-shaped anomalies), the Hazelnut (large object-shaped anomalies), the Tile (large texture-shaped anomalies) and the Wood (small texture-shaped anomalies) dataset, which are made of 480, 501, 347 and 326 images, respectively. All of these datasets are clearly imbalanced with a minority of abnormal images.

To investigate the medical side, four datasets of object and texture-shaped images are used, coming from healthy and unhealthy subjects. First, PCAM (large texture-shaped anomalies) [23] consists in 220,025 histological images. Second, Breast Ultrasound (small texture-shaped anomalies) [24] is made of 789 images. One should mention that for this dataset, many images were manually labeled by experts by highlighting the tumor on the images. Therefore, in order to avoid any bias during training, we removed all those annotated images from the dataset, resulting in a dataset of 654 images. The third dataset is made of 253 Brain MRI images (large object-shaped anomalies) [25]. Finally, the retinal OCT dataset [26] (small and large texture-shaped anomalies) contains 83,600 images of Optical Coherence Tomography. All of these datasets are imbalanced with a minority of normal images, except for PCAM where the abnormal images are slightly minority. Table I provides an overview of the number of images for each dataset.

TABLE I: Sizes of the training and test sets for each dataset, regarding the number of normal and abnormal data. A particular split to guaranty that the test sets are balanced is always chosen, even if the initial datasets are imbalanced.

Dataset	Training set		Test set	
	# Normal	# Abnormal	# Normal	# Abnormal
Wood	236	30	30	30
Tile	222	43	41	41
Hazelnut	396	35	35	35
Screw	301	59	60	60
Breast	67	455	66	66
PCAM	121,998	80,207	8,910	8,910
Retina	13,172	44,084	13,172	13,172
Brain	49	106	49	49

The same data preprocessing steps are performed for the all the datasets, as the industrial and medical images are similar in an object-shaped or texture-shaped point of view. Some of the aforementioned datasets come with different types of anomalies. In this case, a single abnormal class is created by aggregating all the abnormal classes together. Also, the normal and abnormal classes are sometimes imbalanced. Therefore, in order to avoid the use of an imbalanced metric to evaluate the results, a specific split of the dataset is applied to ensure that the test sets are fully balanced. This is obtained by keeping half of the minority class images for the testing set, as well as the same number of randomly picked images from the majority class. All the remaining images are left to the training set¹. Because AD may sometimes be a very imbalanced predictive problem, the majority class is generally overpopulated in the training sets. However, the test sets are perfectly balanced, which allows us to assess the performances with a simple accuracy metric. Furthermore, even if using imbalanced training sets may hurt the performances for most of the supervised machine learning algorithms, the training process of Cycle-GANs is less sensitive to this. The task of cycle-GAN differs from a simple label prediction, and each image gives feedback to the two generators, directly or indirectly. Images are resized to a resolution of 256×256 pixels by using a bicubic interpolation method, except for PCAM where the images remained in their original lower resolution of 96×96 . Data augmentation is also performed so that objects and textures are rotated and flipped along both axes (except for retinal OCT where only flipping along the horizontal axis is pertinent).

B. Network Architecture and Training Procedure

For convenience and practical purposes, the architectures used in this work as well as the training procedures are similar for the different applications, and follow the experimental setup presented in the initial paper on cycle-GANs [1]. The generators are formed by three convolution layers, several

¹An exception is made for the PCAM dataset. Due to its larger size, only 10% of the minority class is taken for testing instead of 50%.

TABLE II: AD accuracy (in %), based on the FID or SSE anomaly scores, with thresholds set to enforce a zero-false-negative constraint (ZFN thr.) or to maximize the AD accuracy (ACC thr.). Best scores are highlighted in bold.

Dataset	Domain	Type	Anomaly Score = FID		Anomaly Score = SSE	
			ZFN thr.	ACC thr.	ZFN thr.	ACC thr.
Wood	Indust.	Texture	93.33	95.00	73.33	83.33
Tile	Indust.	Texture	100.00	100.00	89.02	97.56
Hazelnut	Indust.	Object	100.00	100.00	72.86	94.29
Screw	Indust.	Object	50.83	57.50	52.50	52.50
Breast	Med.	Texture	65.15	93.18	87.12	94.70
PCAM	Med.	Texture	94.41	97.79	82.42	99.89
Retina	Med.	Texture	52.17	96.54	50.29	95.46
Brain	Med.	Object	51.02	62.24	62.24	68.37

residual blocks [27], two fractionally-strided convolution layers and one final convolution layer. We use 9 residual blocks for images resized at 256×256 resolution, and only 6 for the PCAM (96×96 resolution). For the discriminators, we use 70×70 PatchGANs [28]–[30]. All the models are trained through 200 iterations of the Adam optimizer with a learning rate of 2×10^{-4} . An exception is made for PCAM, for which only 40 training iterations are performed due to its large size. A linear learning decay is introduced at the middle of the training. The meta-parameters λ_{cyc} and λ_{ide} are fixed to 10 and 5, respectively. To give an idea of the computation time, training the cycle-GAN on 500 images with a 256×256 resolution for 200 iterations roughly takes 24 hours on a single Nvidia RTX A6000 GPU.

C. Experimental Results

1) *Qualitative Assessments:* The quality of the reconstruction, as well as the highlighting of anomalies are presented in Fig. 3. It shows the original image, the normal (generated) version and their squared pixel difference image, for the selected industrial and medical datasets. These images have been specifically chosen to illustrate different typical cases. However, the following quantitative assessment evaluates the global performances on all the test sets, which are in agreement with the qualitative examples presented here.

2) *Quantitative Assessments:* Table II summarizes the accuracy (for the FID and SSE anomaly scores) under the zero-false-negative constraint (ZFN thr. columns), and, in a more standard way, without this constraint (ACC thr. columns) for all the different datasets. The distributions of the FID anomaly scores for the normal and abnormal test sets are shown in Fig. 4, with the accuracy calculated for the threshold set with the ZFN constraint, or without it.

VI. DISCUSSION AND LIMITATIONS

This section outlines the discussion and the limitations with the use of cycle-GANs, for industrial and medical images AD. We observe from the qualitative results presented in Fig. 3 that the anomaly reconstruction strongly depends on the nature of the image. Indeed, for the textured appearance images (Wood, Tile, Breast Ultrasound, PCAM and OCT Retina), small holes,

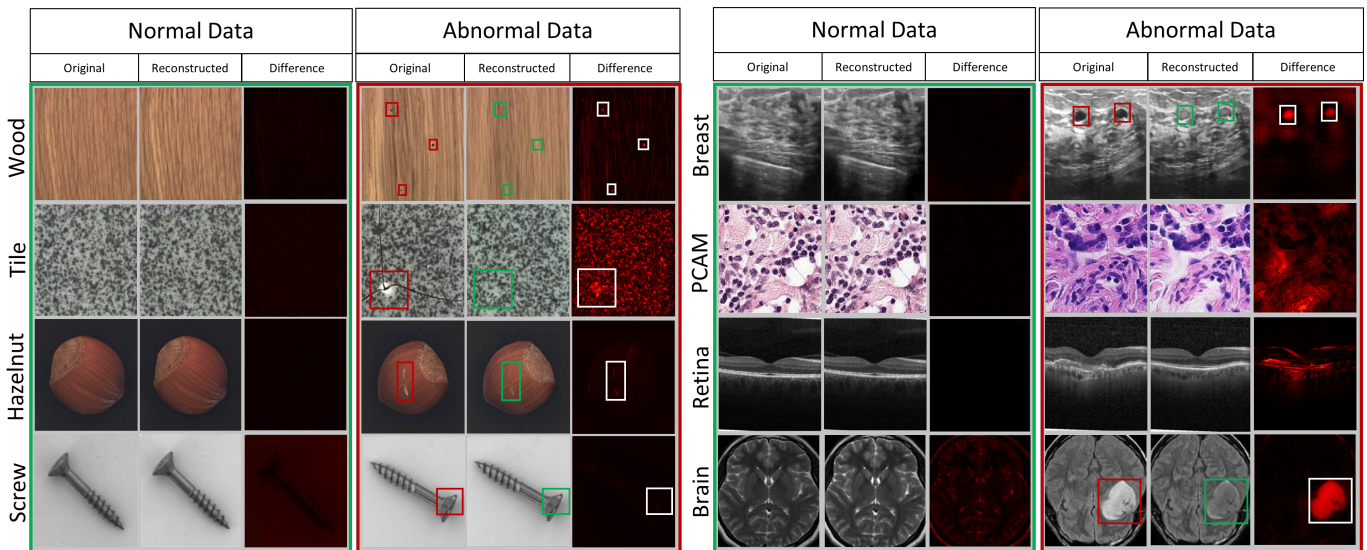


Fig. 3: Industrial (left set) and medical (right set) image examples. For each set of datasets, the left green-framed block presents normal images and the right red-framed block shows abnormal images, with the original image (1st column), the normal version generated (2nd column), and their squared pixel-wise difference image (3rd column). We manually added a small frame around each defect (abnormal images) for the reader’s convenience.

large cracks, blurred areas or colorization contained in the abnormal images (left images of the red-framed blocks) are erased by the estimated normal texture (middle images of the red-framed blocks) and result in a pixel difference image (right images of the red-framed blocks) that faithfully highlights the anomalies. We can note that for complex random textures such as Tile, the model struggles to perfectly restore the normal areas, resulting in a noisy difference image. However, the anomaly can still be localized due to the even greater difficulty in restoring the abnormal area.

Good reconstruction is also observed for the Hazelnut and Brain MRI datasets, where these object-shaped images have large anomalies such as scratches or spots shown at the red-framed block of each set. They are not well erased but attenuated, which provides enough information in the difference image to detect and localize the anomaly. However, for an object-shaped dataset with small defects such as the Screw dataset (red-framed block), the anomaly does not disappear after reconstruction, making it impossible to highlight in the difference image.

We can also observe in the green-framed block (normal images) of each set that the reconstructions (middle images) are more or less identical to the original input (left images), resulting in an almost zero difference image (right images). The model has extracted the features of the normal distributions, and is able to restore normal images without changing the pixels value, thanks to the identity loss. Pixel areas with high discontinuity, as shown in the Brain MRI or Screw images, do not fully follow this observation, resulting in slight differences in the generated image that disturb the anomaly score, and make it difficult to obtain good separable thresholds, in the quantitative step.

For the quantitative assessment, we conclude from Table II that the Tile and the Hazelnut datasets achieve perfect classification on the test set under the FID anomaly score, and with or without the zero false negative constraint (ZFN). Nevertheless, this result could be overestimated due to the absence of many challenging examples that might occur in the real world. As expected from the qualitative assessment, the Wood and the PCAM datasets also perform well, with 93.33% and 94.41% accuracy with the FID anomaly score under the ZFN constraint, better than the 65.15% accuracy for the Breast Ultrasound dataset (even if the standard accuracy is 93.18%) or the 52.17% accuracy for the OCT Retina dataset (even if the standard accuracy is 96.54%). Indeed, the Breast and the OCT Retina datasets contain challenging abnormal images, resulting in a lower anomaly score that explains this high impact of the ZFN constraint on the accuracy. Despite the fact that the method achieves pretty good anomaly localization with the Brain MRI dataset, it has trouble to correctly discriminate the two classes, with an accuracy of 51.02%, due to its noisy normal reconstruction. Finally, the Screw dataset performs also poorly, with a near-random classification accuracy of about 50%, which was expected from the qualitative assessment. Fig. 4 states how normal and abnormal FID anomaly scores distributions overlap for the Screw, the Breast, the OCT Retina and the Brain MRI test sets, resulting in a relative poor accuracy compared to the other datasets. Regarding these results, the method reaches an average accuracy of 97.2% (85.4% with the additional ZFN constraint).

Overall, we can also conclude that the FID anomaly score improves the accuracy for the industrial images, by getting rid of the noisy pixel-by-pixel reconstruction described above. However, this score cannot avoid poor accuracy with an object-

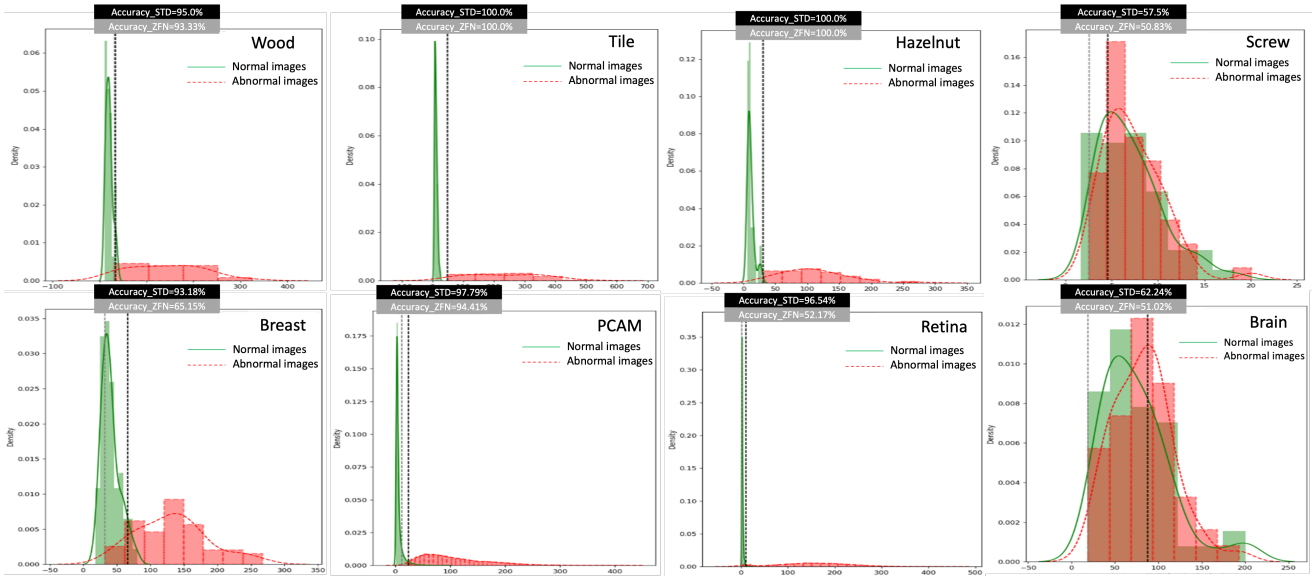


Fig. 4: FID anomaly score distributions of normal (solid-green line and bars) and abnormal (dashed-red line and bars) images for the test datasets, with the threshold value in the ZFN setting (vertical dashed line in grey) or in the ACC setting (vertical dashed line in black).

shaped dataset like Screw, where small anomalies are not well captured by the model and the generated images still show the anomalies. For the Breast Ultrasound and Brain MRI datasets, it appears that for the ZFN constraint, the anomaly score based on SSE leads to better results. This could come from the fact that the Inception V3 model is not pre-trained on many medical-like images, leading in a poor feature extraction. We can also state that the ZFN constraint reduces the accuracy, due to a non-optimal classification threshold, except for the perfect classification of the Hazelnut or Tile datasets. In a way, this is the price to pay for adapting to a real case such as we may encounter in the industrial or medical domain.

To conclude this section, one would like to highlight the advantage of our method compared to state-of-the-art methods. To do so, Table III shows a comparison between a non-exhaustive list of them regarding several criteria. Among those criteria, we checked if other methods use a ZFN constraint, as it can be of fundamental importance in business and medical applications. We also compared the types of data, the different loss functions and whether they use abnormal data during the training phase. It clearly appears that our work covers a wider spectrum of criteria, while being the first GAN-based approach to reintroduce abnormal data during the training loop. As already presented in Section III, GAN is today the most widely used deep learning architecture for AD. Given a large set of normal images, GANs are able to learn a correct representation of them, and generate new samples from it. Afterwards, when feeding the model with abnormal images, differences between the input and reconstructed images may highlight anomalies. However, with such frameworks, the abnormal images are generally not used during the training while they are sometime be easily available (although often in small amount). In such

situations, a cycle-GAN based method can benefit from the use of abnormal data during training in order to refine its representation of normal data.

VII. CONCLUSION

In this work, we propose and characterize for the first time an approach using Cycle-Generative Adversarial Networks (cycle-GANs) for Anomaly Detection (AD) on industrial and medical images. This method allows us to also exploit the abnormal images when at our disposal to refine its representation of normal data, by giving more insights on what is normal or abnormal. Furthermore, thanks to the use of the identity loss, we show that the formalism of cycle-GAN is naturally well-adapted to perform AD. Particular attention has been given to industrial and medical applications, due to the societal impact it may offer, and motivated by the lack of studies for such kind of work in these areas to date. The proposed method differs from previous work by exploiting both normal and abnormal images to learn mappings that can generate new matched data from one domain to another, under a cycle consistency constraint. The mapping of interest for our AD method is the one that can generate normal images. From this perspective, any differences between the test image and its normal (generated) version can be easily identified. Qualitatively, the pixel squared difference image is used to locate abnormal areas, and then quantitatively, an anomaly score is created to indicate whether the image contains abnormal areas, based on a preselected threshold. Ultimately, the method identifies anomalies at the pixel level while the labels are initially at the image level, i.e., without the requirement for tedious annotation at the pixel level.

The achieved results demonstrate that, independent of the application, images with a texture appearance (with continuous

TABLE III: Comparison between a non-exhaustive list of other recent works for AD in images. This table highlights the fact that our cycle-consistent method is complementary to what is proposed in other works. In Cycle-GANs, the reconstruction loss is obtained by the combination of the cycle-consistent loss and the identity loss.

Criteria	Autoencoder [17]	f-AnoGAN [8]	GANomaly [9]	EGBAD [10]	VQGanoDIP [11]	Cycle-GANs <i>Our work</i>
Apply a ZFN constraint	✗	✗	✗	✗	✓	✓
Tested on industrial data	✓	✗	✗	✗	✓	✓
Tested on medical data	✗	✓	✗	✗	✗	✓
Tested on texture-shaped anomalies	✓	✓	✗	✗	✓	✓
Tested on large object-shaped anomalies	✗	✗	✓	✓	✓	✓
Tested on small object-shaped anomalies	✗	✗	✗	✗	✓	✓
Use image with resolution $> 64 \times 64$	✓	✗	✗	✗	✓	✓
Use the GAN loss	✗	✓	✓	✓	✓	✓
Use the reconstruction loss	✓	✓	✓	✗	✓	✓
Use abnormal data for training	✗	✗	✗	✗	✗	✓

pixel value variability, such as colorization or progressive blurred areas) tend to benefit from higher domain change mapping than those with an object appearance (with drastic pixel value changes, such as strong contours or structural shapes). An exception is observed for images of objects with coarse defects where the localization and detection of anomalies always meet expectations. We argue in this work that when both normal and abnormal data are available for training, the use of cycle-GAN architectures should be considered as an approach by the community, mainly when the anomalies are known to be in the form of textures or coarse objects.

New applications may also be explored for future work, such as object segmentation or object counting for industrial and medical fields using the same type of cycle-consistent models. This work is a first step and a proof-of-concept for cycle-GANs in AD for industrial and medical domains.

ACKNOWLEDGMENT

V.D. benefits from the support of the Walloon region with a Ph.D. grant from FRIA (F.R.S.-FNRS). M.E. benefits from the support of the Belgian Walloon region for funding SMART-SENS project which is part of Win²WAL program (agreement 2110108). The authors thank Jérôme Fink and Géraldine Nanfack for their insightful comments and discussions on this paper.

REFERENCES

- [1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *ICCV*. IEEE, 2017.
- [2] “TensorFlow Datasets, a collection of ready-to-use datasets,” <https://www.tensorflow.org/datasets>.
- [3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *NeurIPS*, vol. 27, 2014.
- [4] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*. Springer, 2017, pp. 146–157.
- [5] A. Bougaham, A. Bibal, I. Linden, and B. Frenay, “Ganodip-gan anomaly detection through intermediate patches: a pcba manufacturing case,” in *LIDTA*. PMLR, 2021, pp. 104–117.
- [6] B. R. Kiran, D. M. Thomas, and R. Parakkal, “An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos,” *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

- [7] A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: A survey,” *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.
- [8] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks,” *Medical Image Analysis*, vol. 54, pp. 30–44, May 2019.
- [9] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, “GANomaly: Semi-supervised Anomaly Detection via Adversarial Training,” in *Computer Vision – ACCV 2018*, ser. Lecture Notes in Computer Science, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Springer International Publishing, 2019, pp. 622–637.
- [10] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, “Efficient GAN-Based Anomaly Detection,” May 2019, arXiv:1802.06222 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1802.06222>
- [11] A. Bougaham, M. E. Adoui, I. Linden, and B. Frénay, “Composite Score for Anomaly Detection in Imbalanced Real-World Industrial Dataset,” Nov. 2022, arXiv:2211.15513 [cs]. [Online]. Available: <http://arxiv.org/abs/2211.15513>
- [12] S. Pandey, P. R. Singh, and J. Tian, “An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation,” *Biomedical Signal Processing and Control*, vol. 57, p. 101782, 2020.
- [13] T. Kerepecky, J. Liu, X. W. Ng, D. W. Piston, and U. S. Kamilov, “Dual-cycle: Self-supervised dual-view fluorescence microscopy image reconstruction using cyclegan,” *ArXiv:2209.11729*, 2022.
- [14] D. Dirvanauskas, R. Maskeliūnas, V. Raudonis, R. Damaševičius, and R. Scherer, “Hemigen: human embryo image generator based on generative adversarial networks,” *Sensors*, vol. 19, no. 16, p. 3578, 2019.
- [15] O. Rippel, M. Müller, and D. Merhof, “Gan-based defect synthesis for anomaly detection in fabrics,” in *ETFA*, vol. 1. IEEE, 2020, pp. 534–540.
- [16] G. Zhang, K. Cui, T.-Y. Hung, and S. Lu, “Defect-GAN: High-Fidelity Defect Synthesis for Automated Defect Inspection,” in *WACV*. IEEE, 2021, pp. 2523–2533.
- [17] J. Liu, K. Song, M. Feng, Y. Yan, Z. Tu, and L. Zhu, “Semi-supervised anomaly detection with dual prototypes autoencoder for industrial surface inspection,” *Optics and Lasers in Engineering*, vol. 136, p. 106324, 2021.
- [18] M. Hammami, D. Friboulet, and R. Kéchichian, “Cycle gan-based data augmentation for multi-organ detection in ct images via yolo,” in *ICIP*. IEEE, 2020, pp. 390–393.
- [19] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, “Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks,” *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.
- [20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *NeurIPS*, vol. 30, 2017.
- [21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *CVPR*. IEEE, 2016, pp. 2818–2826.
- [22] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Mvtec ad-a

- comprehensive real-world dataset for unsupervised anomaly detection,” in *CVPR*. IEEE, 2019, pp. 9592–9600.
- [23] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, “Rotation equivariant CNNs for digital pathology,” in *MICCAI*. Springer, 2018, pp. 210–218.
- [24] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, p. 104863, 2020.
- [25] N. Chakrabarty, “Brain MRI images for brain tumor detection, version 1,” Retrieved May 9, 2021 from <https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection>.
- [26] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, . . . , and K. Zhang, “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*. IEEE, 2016, pp. 770–778.
- [28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *CVPR*. IEEE, 2017, pp. 5967–5976.
- [29] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *CVPR*. IEEE, 2017, pp. 105–114.
- [30] C. Li and M. Wand, “Precomputed real-time texture synthesis with markovian generative adversarial networks,” in *ECCV*. Springer, 2016, pp. 702–716.