Improved anomaly detection by training an autoencoder with skip connections on images corrupted with Stain-shaped noise

Anne-Sophie Collin and Christophe De Vleeschouwer
ICTEAM Institute, UCLouvain
Louvain-La-Neuve, Belgium
{anne-sophie.collin, christophe.devleeschouwer}@uclouvain.be

***Abstract*—In industrial vision, the anomaly detection problem can be addressed with an autoencoder trained to map an arbitrary image, i.e. with or without any defect, to a clean image, i.e. without any defect. In this approach, anomaly detection relies conventionally on the reconstruction residual or, alternatively, on the reconstruction uncertainty. To improve the sharpness of the reconstruction, we consider an autoencoder architecture with skip connections. In the common scenario where only clean images are available for training, we propose to corrupt them with a synthetic noise model to prevent the convergence of the network towards the identity mapping, and introduce an original Stain noise model for that purpose. We show that this model favors the reconstruction of clean images from arbitrary real-world images, regardless of the actual defects appearance. In addition to demonstrating the relevance of our approach, our validation provides the first consistent assessment of reconstruction-based methods, by comparing their performance over the MVTec AD dataset [1], both for pixel- and image-wise anomaly detection. Our implementation is available at https://github.com/anncollin/AnomalyDetection-Keras.**

I. Introduction

Anomaly detection can be defined as the task of identifying all diverging samples that does not belong to the distribution of regular, also named clean, data. When considering the specific application of the visual inspection of production lines, we are interested in detecting all defective samples occurring due to an unexpected behavior of the manufacturing process. This anomaly detection task could be formulated as a supervised learning problem. Such an approach uses both clean and defective examples to learn how to distinguish these two classes or even to refine the classification of defective samples into a variety of subclasses. However, the scarcity and variability of the defective samples make the data collection challenging and frequently produce unbalanced datasets [2]. To circumvent the above-mentioned issues, anomaly detection is often formulated as an unsupervised learning task. This formulation makes it possible to either solve the detection problem itself or to ease the data collection process required by a supervised approach.

The unsupervised anomaly detection framework considered in this work is depicted in Figure 1. It builds on the training of an autoencoder to project an arbitrary image onto the clean distribution of images (blue block). The training set is constituted exclusively of clean images. Then, defective

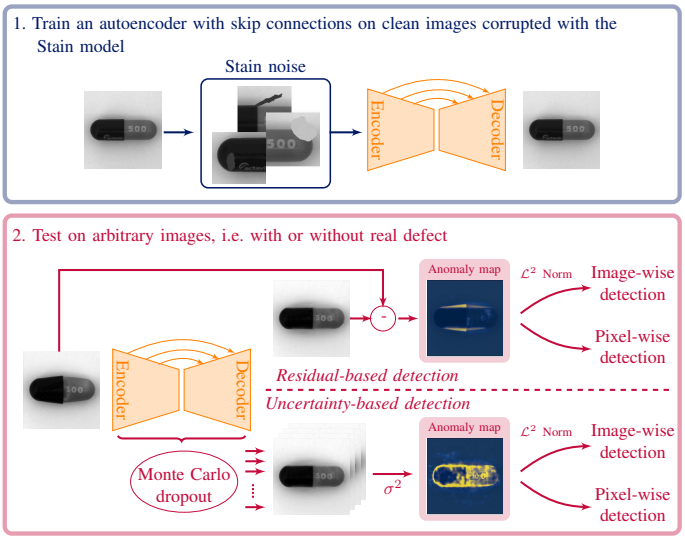


Fig. 1. We improve the quality of the reconstructed images by training an autoencoder with skip connections over corrupted images. *1. Blue block.* Corrupting the training images with our Stain noise model avoids the convergence of the network towards an unwanted identity operator. *2. Red block.* The two anomaly detection strategies. In the upper part, the anomaly map is generated by subtracting the input image from its reconstruction. In the lower part, the anomaly map is estimated by the variance between 30 reconstructions inferred with Monte Carlo dropout (MCDropout) [4]. It relies on the hypothesis that structures that are not seen during training (defective areas) correlate with higher reconstruction uncertainty.

structures can be inferred from the reconstruction (red block), following a traditional approach based on the residual [2], or even from an estimation of the prediction uncertainty [3].

In conventional reconstruction-based approach, an autoencoder is trained on clean images only to perform an identity mapping. The bottleneck forces the network to learn a compressed representation of the training data that is expected to regularize the reconstruction towards the normal class. In the literature, the use of the Mean Squared Error (MSE) loss to train an hourglass CNN, without skip connections, has been criticized for its trend to produce blurry output images [5], [6]. Since the anomaly detection is based on the reconstruction residual, this behavior is detrimental

because it alters the clean structures of an image as well as the defective ones.
A lot of effort has been made to improve the quality of the reconstructed images by the introduction of new loss functions. In this spirit, unsupervised methods based on Generative Adversarial Networks (GANs) have emerged [7], [8], [9], [10], [11]. If GANs are known for their ability to produce realistic high-quality synthetic images [12], they have major drawbacks. Usually, GANs are difficult to train due to their trend to converge towards mode collapse [13]. In the context of anomaly detection, some GAN-based solutions fail to exclude defective samples from the generative distribution [10] and require an extra optimization step in the latent space during inference [8]. Performances of AnoGAN [8] over the MVTec AD dataset have been reported by Bergmann et al. [1]. Those are significantly lower than the method proposed in this work.
Also, the use of a loss based on the Structural SIMilarity index has been considered for image generation, motivated by its ability to produce well looking images from a human perceptual perspective [5], [14]. The SSIM have shown some improvement over the MSE loss for the training of an autoencoder in the context of anomaly detection. However, the SSIM loss formulation does not generalize to color images and is parametric. Traditionally, these hyper-parameters are tuned based on a validation set. However, in a real-life scenarios of anomaly detection, samples with real defects are usually not available. For this reason, our paper focus on the MSE rather than on the parametric SSIM.

With the objective of building our method on the MSE loss for its simplicity and widespread usage, we propose a new non-parametric approach that addresses the above-mentioned issues. To enhance the sharpness of the reconstruction, we consider an autoencoder equipped with skip connections, which allow the information to bypass the bottleneck. In order to prevent systematic transmission of the image structures through these links, the network is trained to reconstruct a clean image out of a corrupted version of it. As discussed later, the methodology used to corrupt the training images has a huge impact on the overall performances. We introduce a new synthetic model, named Stain, that adds an irregular elliptic structure of variable color and size to the input image. Despite its simplicity, the Stain model is by far the best performing compared to the scene-specific corruption investigated in a previous study [15]. Our Stain model has the double advantage of performing consistently better, while being independent of the image content.

In Section II we provide an overview of previous reconstruction-based methods addressing the anomaly detection problem. Details of our method, including network architecture and the Stain noise model description, are provided in Section III. In Section IV we provide a comparative study of residual- and uncertainty-based anomaly detection strategies, both at the image and pixel level. To the best of our knowledge, our work is the first to provide a fair comparison (using the same dataset, comparable network architectures, covering a large variety of use cases) between the various detection strategies proposed in the recent literature. This extensive comparative study demonstrate the benefit of our proposed framework, combining skip connections and our novel corruption model.

II. Related Work

Anomaly detection is a long-standing problem that has been considered in a variety of fields [2], [16] and the reconstruction-based approach is one popular way to address the issue. In comparison to other methods for which the detection of abnormal samples is performed in another domain than the image [17], [18], [19], reconstruction-based approaches offer the opportunity to identify the pixels that lead to the rejection of the image from the normal class.

Conventional reconstruction-based methods infer anomaly based on the reconstruction error between an arbitrary input and its reconstructed version. It assumes that clean structures are perfectly conserved while defective ones are replaced by clean content. However, when a defect contrasts poorly with its surroundings, replacing abnormal structures with clean content does not lead to a sufficiently high reconstruction error. In such cases, this methodology reaches the limit of its underlying assumptions. A previous study [3] detected anomalies by quantifying the prediction uncertainty with MCDropout [4] instead of the reconstruction residual.

To obtain a clean reconstruction out of an arbitrary image, an autoencoder, without skip connections, is generally trained to perform an image-to-image identity mapping with clean data only under the minimization of the MSE. This loss has the disadvantage of promoting blurry reconstructions, resulting in higher residuals for clean structures.
To improve the sharpness of the reconstruction, Bergmann et al. proposed a loss derived from the SSIM index [6]. However, the SSIM imposes to consider grayscale images and depends on multiple hyper-parameters, thereby hampering the reproducibility of the results.
Also, GANs have been considered to sample the clean distribution of images [7], [8], [9], [10], [11]. Unfortunately, GANs are challenging to train due to their trend to converge towards mode collapse. Furthermore, the difficulty to exclude abnormal samples from the generative distribution penalizes performances [10]. Finally, some GAN-based methods require an optimization process during inference to find the latent space vector producing the most similar image between an arbitrary query and an output image belonging to the generative distribution [8].

Excluding defective structures from the distribution of generated images is a recurrent problem in anomaly detection. It is usually expected that the compression induced by the bottleneck is sufficient to regularize the reconstruction so

that it lies on the clean images manifold. In practice, the autoencoder is not explicitly constrained to not reproduce abnormal content and often reconstructs defective structures. As an extension to traditional autoencoder-based methods studied in this work, a recent method proposed to mitigate this issue by iteratively projecting the arbitrary input towards the clean distribution of images. The projection is constrained to be similar, in the sense of the $\mathcal{L}^1$ norm, to the initial input [20]. Instead of performing this optimization in the latent space as made with AnoGAN [8], they propose to find an optimal clean input image. If this practice enhances the sharpness of the reconstruction, the optimization step is resource consuming.
Also, the reconstruction task can be formulated as an image completion problem [21], [22]. To make the inference and training phases consistent, it is assumed that the defects are entirely contained in the mask during inference, which limits the practical usage of the method. Mei et al. [23] also proposed to use a denoising autoencoder to reconstruct training images corrupted with salt-and-pepper noise. However they did not discuss the gain brought by this modification, and only considered it for an hourglass CNN, without skip connections.

The methodology proposed in this work presents a simple approach to enhance the sharpness of the reconstruction. The skip connections allow the preservation of high frequency information by bypassing the bottleneck. However, we show that this practice penalizes anomaly detection when the model is trained to perform identity mapping on uncorrupted clean images. Nevertheless, the introduction of an original noise model allows to significantly improve the anomaly detection accuracy for the skipped architecture, which eventually outperforms the conventional one in many real-life cases.

III. Methods

Our method performs anomaly detection based on the regularized reconstruction predicted by an autoencoder. This section presents the different components of our approach, ranging from the training of the autoencoder to the strategies considered to detect defects based on the reconstruction residual or the reconstruction uncertainty.

A. Model configuration

The reconstruction is based on a convolutional neural network. Our architecture, referred to as **Autoencoder with Skip connections (AESc)** and shown in Figure 2, is a variant of U-Net [24]. AESc takes input images of size 256×256 and projects them onto a latent space of $4 \times 4 \times 512$ dimension. The projection towards the lower dimensional space is performed by 6 consecutive convolutional layers strided by a factor 2. The back projection is performed by 6 layers of convolution followed by an upsampling operator of factor 2. All convolutions have a 5×5 kernel. Unlike the original U-Net version, our skip connections perform an addition, not a concatenation, of feature maps of the encoder to the decoder.
For the sake of comparison, we also consider the **Autoencoder (AE)** network which follows exactly the same architecture but from which we removed the skip connections.

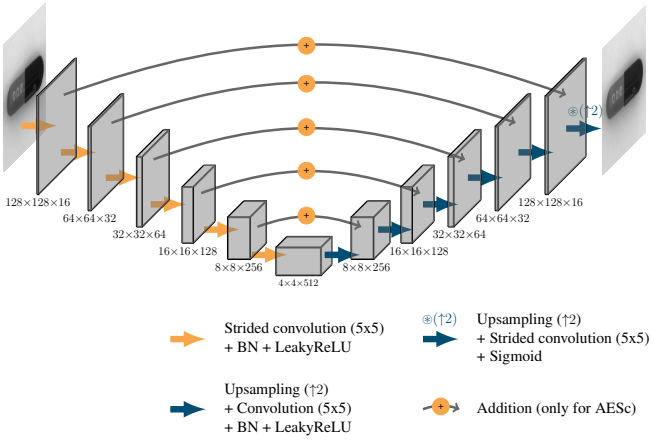


Fig. 2. AESc architecture performing the projection of an arbitrary 256×256 image towards the distribution of clean images with the same dimension. Note that the AE architecture shares the same specifications with the exception of the skip connections that have been removed.

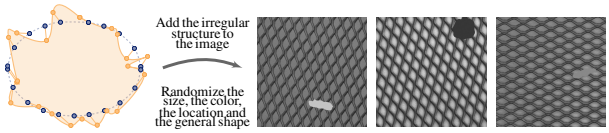


Fig. 3. The Stain noise model is a cubic interpolation between 20 points (orange dots), arranged in ascending order of polar coordinates, located around the border of an ellipse of variable size (blue line). The axes of the ellipse are comprised between 1 and 12% of the smallest image dimension and its eccentricity is randomly initialized.

B. Corruption model

Ideally, the autoencoder should preserve clean structures while modifying those that are not. Due to the impossibility of collecting pairs of clean and defective versions of the same sample, we propose to introduce synthetic corruption during training to explicitly constrain the autoencoder to remove this additive noise. Our **Stain** model, illustrated in Figure 1 and explained in Figure 3, corrupts images by adding a structure whose color is randomly selected in the grayscale range and whose shape is an ellipse with irregular edges.

The intuition behind the definition of this noise model is that occluding large area of the training images is a data augmentation procedure that helps to improve the network training [25], [26]. Due to the skip connections in our network architecture, this form of data augmentation is essential to avoid the convergence of the model towards the identity operator. However, we noticed that the use of regular shapes, like a conventional ellipse, leads to overfitting to this additive noise structure, as also pointed out in a context of inpainting with rectangular holes [27].

C. Anomaly detection strategies

We compare two approaches to obtain the anomaly map representing the likelihood that a pixel is abnormal. On the one hand, the **residual-based** approach evaluates the abnormality by measuring the absolute difference between the input image $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$. On the other hand, the **uncertainty-based** approach relies on the intuition that structures that are not seen during training, i.e. the anomalies, will correlate with higher uncertainties, as estimated by the variance between 30 output images inferred with the MCDropout technique. Our experiments revealed that more accurate detection is obtained by applying an increasing level of dropout for deepest layers. More specifically, the dropout levels are [0, 0, 10, 20, 30, 40] percent for layers ranging from the highest spatial resolution to the lowest.

Out of the anomaly map, it is either possible to classify the entire image as clean/defective or to classify each pixel as belonging to a clean/defective structure. In the first case, referred to as **image-wise detection**, it is common to compute the $\mathcal{L}^p$ norm of the anomaly map given by

$$\mathcal{L}^p(\mathbf{x}, \hat{\mathbf{x}}) = \left(\sum_{i=0}^{m} \sum_{j=0}^{n} |\mathbf{x}_{i,j} - \hat{\mathbf{x}}_{i,j}|^p \right)^{1/p} \quad (1)$$

with $\mathbf{x}_{i,j}$ denoting the pixel belonging to the i^{th} row and the j^{th} column of the image $\mathbf{x}$ of size $m \times n$. Based on our experiments, we present results obtained for $p = 2$ since they achieve the most stable accuracy values across the experiments. Hence, all images for which the $\mathcal{L}^2$ norm of the abnormality map exceeds a chosen threshold are considered as defective. In the second case, referred to as **pixel-wise detection**, the threshold is applied directly on each pixel value of the anomaly map.

To perform image-wise or pixel-wise anomaly detection, a threshold has to be determined. Since this threshold value is highly dependent on the application, we present the performances in terms of Area Under the receiver operating characteristic Curve (AUC), obtained by varying over the full range of threshold values.

IV. Results

Experiments have been conducted on grayscale images of the MVTec AD dataset [1], containing 5 categories of textures and 10 categories of objects. In this dataset, defects are real and have various appearance. Their location is defined with a binary segmentation mask. All images have been scaled to a 256×256 size. Anomaly detection is performed at this resolution.

A. AESc + Stain: qualitative and quantitative analysis

In this section, we compare qualitatively and quantitatively the results obtained with our AESc + Stain model for both image- and pixel-wise detection. We focus this first analysis on the residual-based detection approach to emphasize the benefits of adding skip connections to the AE architecture. The comparison of the results obtained with residual- versus uncertainty-based strategies is discussed later in Section IV-C.

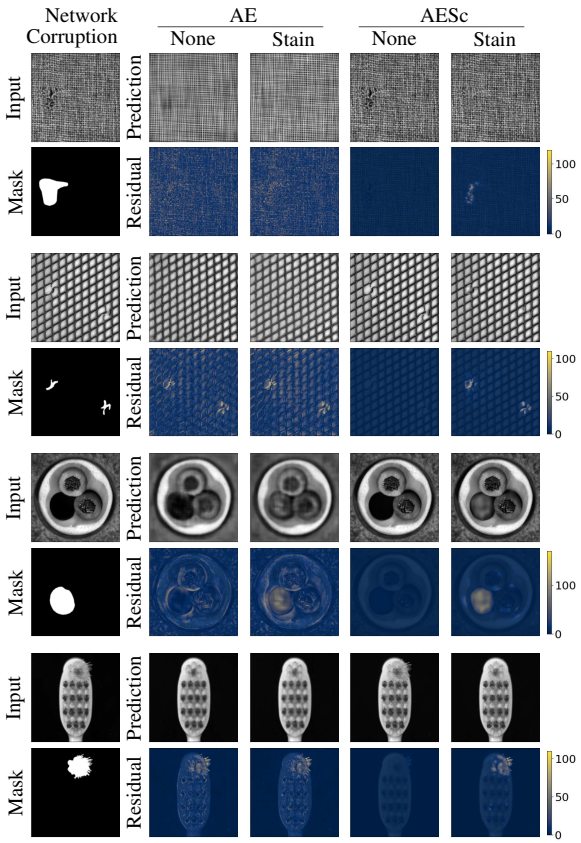


Fig. 4. Predictions obtained with the AESc and AE networks trained with and without our Stain noise model. Two defective textures are considered, namely a carpet (first sample) and a grid (second sample), as well as two defective objects, namely a cable (third sample) and a toothbrush (fourth sample). First column show the image fed in the networks and the mask locating the defect. Odd rows show the reconstructed images and even rows show the anomaly maps obtained with the residual-based strategy.

Qualitatively, Figure 4 reveals the general trends of the AE and AESc models trained with and without the Stain noise corruption. On the one hand, the AE network produces blurry reconstructions as depicted by the overall higher residual intensities. If the global structure of the object images (cable and toothbrush) are properly reconstructed, the AE network struggles to infer the finer details of the texture images (carpet sample). On the other hand, the AESc model shows finer reconstruction of the image details depicted by a nearly zero residual over the clean areas of the images. However, when ASEc is trained without corruption, the model converges towards an identity operator, as revealed by the close-to-zero residuals of defective structures. The corruption of the training images with the Stain model alleviates this unwanted behavior by leading to high reconstruction residuals in defective areas while simultaneously keeping low reconstruction residuals in clean structures.

TABLE I
IMAGE-WISE DETECTION AUC OBTAINED WITH THE RESIDUAL- AND UNCERTAINTY-BASED DETECTION METHODS[a].

		Uncertainty				Residual				
	Network	AE		AESc		AE		AESc		ITAE [15]
	Corruption	None	Stain	None	Stain	None	Stain	None	Stain	
Textures	Carpet	0.41	0.30	0.44	0.80	0.43	0.43	0.48	**0.89**	0.71
	Grid	0.69	0.66	0.12	**0.97**	0.80	0.84	0.52	**0.97**	0.88
	Leather	0.86	0.57	0.88	0.72	0.45	0.54	0.56	**0.89**	0.87
	Tile	0.73	0.50	0.72	0.95	0.49	0.57	0.88	**0.99**	0.74
	Wood	0.87	0.86	0.78	0.78	0.92	0.94	0.92	**0.95**	0.92
	Mean[b]	0.71	0.58	0.59	0.84	0.62	0.66	0.67	**0.94**	0.82
Objects	Bottle	0.72	0.41	0.71	0.82	**0.98**	0.97	0.77	**0.98**	0.94
	Cable	0.64	0.48	0.52	0.87	0.70	0.77	0.55	**0.89**	0.83
	Capsule	0.55	0.49	0.44	0.71	**0.74**	0.64	0.60	**0.74**	0.68
	Hazelnut	0.83	0.60	0.68	0.90	0.90	0.88	0.85	**0.94**	0.86
	Metal Nut	0.38	0.33	0.41	0.62	0.57	0.59	0.24	**0.73**	0.67
	Pill	0.63	0.48	0.55	0.62	0.76	0.76	0.70	**0.84**	0.79
	Screw	0.45	0.77	0.13	0.80	0.68	0.60	0.30	0.74	**1.00**
	Toothbrush	0.36	0.44	0.51	0.99	0.93	0.96	0.78	**1.00**	**1.00**
	Transistor	0.67	0.59	0.55	0.90	0.84	0.85	0.46	**0.91**	0.84
	Zipper	0.44	0.41	0.70	0.93	0.90	0.88	0.72	**0.94**	0.80
	Mean[c]	0.57	0.50	0.52	0.82	0.80	0.79	0.60	**0.87**	0.84
	Global mean[d]	0.62	0.53	0.54	0.83	0.74	0.75	0.62	**0.89**	0.84

[a] For each row, the best performing approach is highlighted in boldface and the second best is underlined.
[b] Mean AUC obtained over the classes of images belonging to the texture categories.
[c] Mean AUC obtained over the classes of images belonging to the object categories.
[d] Mean AUC obtained over the entire dataset.

TABLE II
PIXEL-WISE DETECTION AUC OBTAINED WITH THE RESIDUAL- AND UNCERTAINTY-BASED DETECTION METHODS[a].

		Uncertainty				Residual				
	Network	AE		AESc		AE		AESc		AE_{L2} [1]
	Corruption	None	Stain	None	Stain	None	Stain	None	Stain	
Textures	Carpet	0.55	0.54	0.43	**0.91**	0.57	0.62	0.52	0.79	0.59
	Grid	0.52	0.49	0.50	**0.95**	0.81	0.82	0.57	0.89	0.90
	Leather	0.86	0.52	0.58	0.87	0.79	0.82	0.71	**0.95**	0.75
	Tile	0.54	0.50	0.53	**0.79**	0.45	0.54	0.62	0.74	0.51
	Wood	0.61	0.48	0.51	**0.84**	0.64	0.71	0.65	**0.84**	0.73
	Mean[b]	0.62	0.51	0.51	**0.87**	0.65	0.70	0.61	0.84	0.70
Objects	Bottle	0.68	0.63	0.64	**0.88**	0.85	**0.88**	0.47	0.84	0.86
	Cable	0.54	0.70	0.66	0.84	0.62	0.83	0.72	0.85	**0.86**
	Capsule	0.92	0.89	0.65	**0.93**	0.87	0.87	0.63	0.83	0.88
	Hazelnut	**0.95**	0.91	0.60	0.89	0.92	0.93	0.79	0.88	**0.95**
	Metal Nut	0.79	0.73	0.50	0.62	0.82	0.84	0.52	0.57	**0.86**
	Pill	0.82	0.82	0.61	**0.85**	0.81	0.81	0.64	0.74	**0.85**
	Screw	0.94	0.94	0.61	0.95	0.93	0.93	0.72	0.86	**0.96**
	Toothbrush	0.84	0.83	0.79	**0.93**	0.92	**0.93**	0.73	**0.93**	**0.93**
	Transistor	0.79	0.64	0.51	0.78	0.79	0.82	0.56	0.80	**0.86**
	Zipper	0.78	0.77	0.60	**0.90**	0.73	0.75	0.60	0.78	0.77
	Mean[c]	0.81	0.79	0.62	0.86	0.83	0.86	0.64	0.81	**0.88**
	Global mean[d]	0.74	0.69	0.58	**0.86**	0.77	0.81	0.63	0.82	0.82

[a] For each row, the best performing approach is highlighted in boldface and the second best is underlined.
[b] Mean AUC obtained over the classes of images belonging to the texture categories.
[c] Mean AUC obtained over the classes of images belonging to the object categories.
[d] Mean AUC obtained over the entire dataset.

Quantitatively, the image-wise detection performances obtained with the AESc and AE networks trained with and without our Stain noise model are presented in Table I. The last column provides a comparison with the ITEA method, introduced by Huang et al. [15]. ITAE is also a reconstruction-based approach which relies on an autoencoder with skip connections trained with images corrupted by random rotations and a graying operator (averaging of pixel value along the channel dimension) selected based on prior knowledge about the task.

This table highlights the superiority of our AESc + Stain noise model to solve the image-wise anomaly detection. The improvement brought by adding skip connections to an autoencoder trained with corrupted images is even more important for texture images than for object images. We also observe that, if the highest accuracy is consistently obtained with the residual-based approach, the uncertainty-based decision derived from the AESc + Stain model generally provides the second best (underlined in Table I) performances among tested networks, attesting the quality of the AESc + Stain model for image-based decision.

Table II presents the pixel-wise detection performances obtained with our approaches and compares them with the method reported in [1], referred to as AE_{L2}. This residual-based method relies on an autoencoder without skip connections, and provides SoA performance in the pixel-wise detection scenario. Similarly to our AE model, AE_{L2} is trained to minimize the MSE of the reconstruction of images that are not corrupted with synthetic noise. AE_{L2} however differs from our AE model in several aspects, including a different network architecture, data augmentation, patch-based inference for the texture images, and anomaly map post-processing with mathematical morphology. Despite our efforts, in absence of public code, we have been unable to reproduce the results presented in [1]. Hence, our table just copy the results from [1]. For fair comparison between AE and AESc + Stain, the table also provides the results obtained with our AE, since our AE and AESc + Stain models adopt the same architecture (up to the skip connections) and the same training procedure.

In the residual-based detection strategy, our AESc + Stain method obtains similar performances as the AE_{L2} approach when averaged over all the image categories of the MVTec AD dataset. However, as already pointed in the image-wise detection scenario, we notice that AESc + Stain performs better with texture images and worse with object images. Regarding the decision strategy, we observe an opposite trend than the one encountered for image-wise detection: the uncertainty-based approach performs a bit better than the residual-based strategy when it comes to pixel-wise decisions. This difference is further investigated in the next section.

B. Residual- vs. uncertainty-based detection strategies

Figure 5 provides a visual comparison between residual- and uncertainty-based strategies. Globally, we observe that the reconstruction residual mostly correlates with the uncertainty. However, the uncertainty indicator is usually

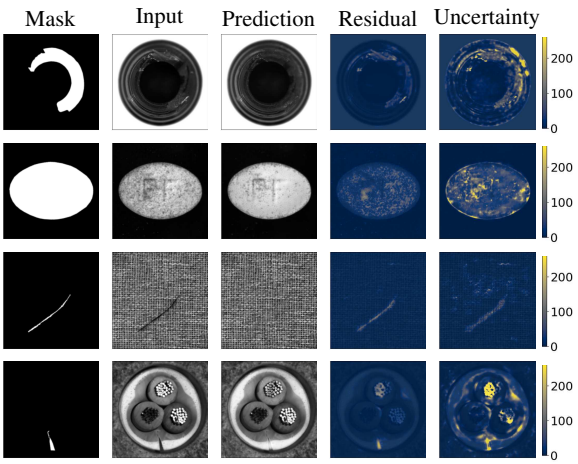


Fig. 5. Predictions obtained with the AESc network trained with our Stain noise model. One defective texture is considered, namely a carpet (third row) as well as three defective objects, namely a bottle (first row), a pill (second row) and a cable (fourth row). From left to right, columns represent the ground-truth, the image fed to the network, the prediction (without MCDropout), the reconstruction residual and the reconstruction uncertainty.

more widespread. This behavior can sometimes lead to a better coverage of the defective structures (bottle and pill) or increase the number of false positive pixels that are detected (carpet and cable).

One important observation concerns the relationship between the detection of a defective structure and its contrast with its surroundings. In the residual-based approach, regions of an image are considered as defective if their reconstruction error exceeds a threshold. In the proposed formulation, the network is explicitly constrained to replace synthetic defective structures with clean content. No constraint is introduced regarding the contrast of the reconstructed structure and its surroundings. Hence, defects that are poorly contrasted lead to small residual intensities. On the contrary, the intensity of the uncertainty indicator does not depend on the contrast between a structure with the surroundings. For low contrast defects, it enhances their detection as illustrated (bottle and pill). On the contrary, it can deteriorate the location of high contrast defects for which the residual map is an appropriate anomaly indicator (carpet and cable). In theses cases, the sharp prediction obtained with the residual-based approach is preferred over the uncertainty-based one.

As reported in Section IV-A, we observe that the uncertainty-based detection perform generally worse than the residual-based approach for image-wise detection. We explain this drop of performance by an increase of the intensities of the uncertainty maps inferred from the clean images belonging to the test set. As the image-wise detection is based on the $\mathcal{L}^2$ norm of the anomaly map, the lowest the anomaly maps of clean images, the better the detection of defective images. For image-wise detection, the performances are less sensitive to the optimal coverage of the defective area as long as the overall intensity of the clean anomaly maps is low.

On the contrary, the uncertainty-based strategy improves the pixel-wise detection of the AESc + Stain model. For this use case, a better coverage of the defective structure is crucial. As previously mentioned, AESc + Stain model used usually leads to reconstruction residual constituted of sporadic spots and misses low contrast defects. The uncertainty-based strategy compensates these two issues.

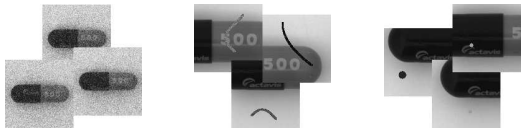
C. Comparative study of corruption models

Up to now, we considered only the Stain noise model to corrupt training data. In this comparative study we consider other noise models to confirm the relevance of our previous approach over other types of corruption that could have been considered. We provide here a comparison with three other synthetic noise models represented in Figure 6:

a- Gaussian noise. Corrupt by adding white noise applied uniformly over the entire image. For normalized intensities between 0 and 1, a corrupted pixel value x', corresponding to an initial pixel value x, is the realization of a random variable given by a normal distribution of mean x and variance σ^2 in the set: $[0.1, 0.2, 0.4, 0.8]$.

b- Scratch. Corrupt by adding one curve connecting two points whose coordinates are randomly chosen in the image and whose color is randomly selected in the gray scale range. The curve can follow a straight line, a sinusoidal wave or the path of a square root function.

c- Drops. Corrupt by adding 10 droplets whose color are randomly selected in the gray scale range and whose shape are circular with a random diameter (chosen between 1 and 2% of the smallest image dimension). The droplets partially overlap.



(a) Gaussian noise. (b) Scratch. (c) Drops.

Fig. 6. Illustration of the Gaussian noise, Scratch and Drops models. The original clean image is the one presented in Figure 1.

In addition, we have also considered the possibility to corrupt the training images with a combination of several models. We propose two hybrid models:

d- Mix1. This configuration corrupts training images with a combination of the Stain, Scratch and Drops models. We fix that 60% of the training images are corrupted with the Stain model while the remaining 40% are corrupted with the Scratch and Drops models in equal proportions.

e- Mix2. This configuration corrupts training images with a combination of the Stain and the Gaussian noise models. We fix that 60% of the training images are corrupted with the Stain model while the remaining 40% are corrupted with the Gaussian noise model.

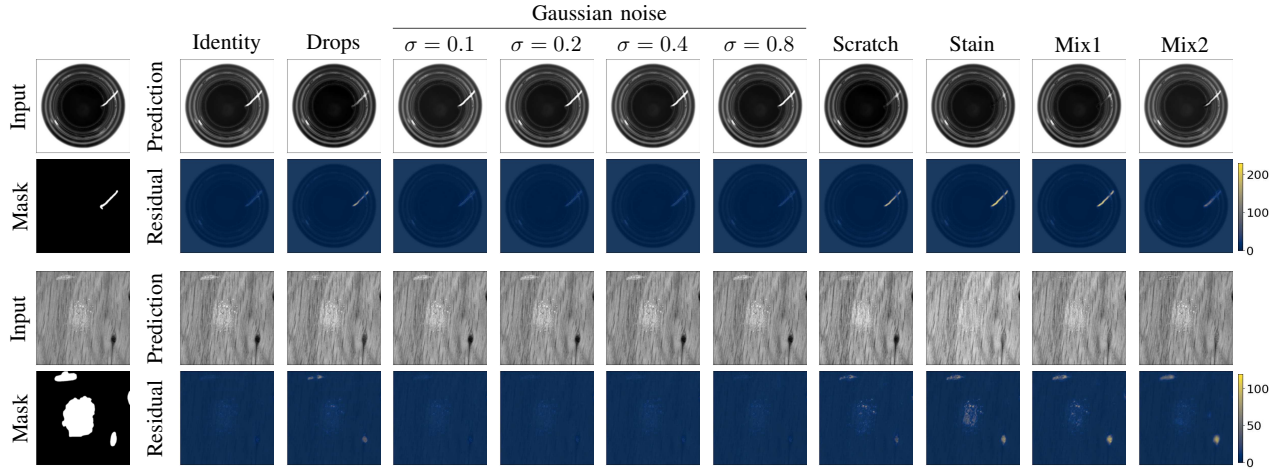


Fig. 7. Reconstructions obtained with the AESc network trained with different noise models. We consider here one defective object, namely a bottle (first sample) and a defective texture, namely a wood (third sample). Rows and columns are defined as in Figure 4.

Figure 7 allows the comparison of the newly introduced noise models with the Stain one over similar samples. First, these examples illustrate the convergence of the model towards the identity mapping when the Gaussian noise model is used as synthetic corruption. An analysis of the results obtained over the entire dataset reveals that the AESc + Gaussian noise model does almost not differ from the AESc network trained with unaltered images.

Compared to the the Gaussian noise, other models introduced before improve the identification of defective areas in the images. This is reflected by higher intensities of the reconstruction residual in the defective areas and close-to-zero reconstruction residual in the clean areas. With the exception of the Gaussian noise model, the Scratch model is the most conservative, among those considered, in the sense that most of the structures of the input images tend to be reconstructed identically. This practice increases the number of false negative. Also, the Drops model restricts the structures detected as defective to sporadic spots. Finally, the three models based on the Stain noise (Stain, Mix1 and Mix2) provide the residuals that correlate the most with the segmentation mask.

Generally, models based on the Stain noise (Stain, Mix1 and Mix2) lead to the most relevant reconstruction for anomaly detection, i.e. lower residual intensities in clean areas and higher residual intensities in defective areas. More surprisingly, this statement remains true even if the actual defect looks more similar to the Scratch model than the Stain noise (bottle sample in Figure 7). We recall that defects contained in the MVTec AD dataset are real observations of an anomaly. This reflects that models trained with synthetic corruption models that look similar to real ones do not necessarily generalize well to real defects

Table III quantifies the impact of the synthetic noise model on the performances of the ASEc network to solve the image-wise detection task with a residual-based approach. The AESc + Stain configuration is the best performing in all use cases when considering the mean performances that are obtained over the entire dataset, as revealed by the previous qualitative study. The two hybrid models (Mix1 and Mix2) lead usually to sightly lower performances than those obtained with the Stain model. Those observations attest that the Stain model is superior to others and justify the choice of the Stain noise as our newly introduced approach to corrupt the training images with synthetic noise.

TABLE III
IMAGE-WISE AUC OBTAINED WITH THE AESC NETWORK TRAINED WITH DIFFERENT NOISE MODELS WITH THE RESIDUAL-BASED PROBLEM FORMULATION[a].

	Corruption	None	Drops	Gaussian noise (σ)				Scratch	Stain	Mix1	Mix2
				0.1	0.2	0.4	0.8				
Textures	Carpet	0.48	0.87	0.52	0.46	0.51	0.53	0.63	**0.89**	0.84	0.84
	Grid	0.52	0.94	0.55	0.69	0.59	0.72	0.79	**0.97**	0.91	0.96
	Leather	0.56	0.87	0.72	0.71	0.74	0.71	0.77	**0.89**	0.88	**0.89**
	Tile	0.88	0.94	0.94	0.92	0.90	0.92	0.95	**0.99**	0.98	0.96
	Wood	0.92	**0.99**	0.89	0.90	0.91	0.85	0.96	0.95	0.94	0.79
	Mean[b]	0.67	0.92	0.72	0.74	0.73	0.75	0.82	**0.94**	0.91	0.89
Objets	Bottle	0.77	**0.99**	0.82	0.85	0.81	0.75	0.91	0.98	0.98	0.97
	Cable	0.55	0.60	0.58	0.53	0.49	0.46	0.60	0.89	0.87	**0.90**
	Capsule	0.60	0.71	0.58	0.68	0.57	0.59	0.66	**0.74**	**0.74**	0.53
	Hazelnut	0.85	**0.98**	0.75	0.73	0.92	0.73	0.96	0.94	0.93	0.81
	Metal Nut	0.24	0.54	0.32	0.27	0.28	0.24	0.44	0.73	0.71	**0.86**
	Pill	0.70	0.79	0.69	0.71	0.73	0.68	0.78	**0.84**	0.77	0.78
	Screw	0.30	0.46	0.91	**0.99**	0.78	0.65	0.71	0.74	0.22	0.72
	Toothbrush	0.78	**1.00**	0.99	0.98	0.79	0.82	0.87	**1.00**	**1.00**	**1.00**
	Transistor	0.46	0.83	0.55	0.49	0.48	0.50	0.68	0.91	**0.92**	**0.92**
	Zipper	0.72	0.93	0.66	0.63	0.69	0.58	0.79	0.94	0.90	**0.98**
	Mean[c]	0.60	0.78	0.68	0.69	0.65	0.60	0.74	**0.87**	0.80	0.85
	Global mean[d]	0.62	0.83	0.70	0.70	0.68	0.65	0.77	**0.89**	0.84	0.86

[a] For each row, the best performing approach is highlighted in boldface and the second best is underlined.
[b] Mean AUC obtained over the classes of images belonging to the texture categories.
[c] Mean AUC obtained over the classes of images belonging to the object categories.
[d] Mean AUC obtained over the entire dataset.

V. Conclusion

In this work, we considered an anomaly detection method based on the reconstruction of a clean image from any arbitrary image. It builds on convolutional autoencoder and relies on the reconstruction residual or the prediction uncertainty, estimated with the Monte Carlo dropout technique, to detect anomalies. We demonstrated the benefits of considering an autoencoder architecture equipped with skip connections, as long as the training images are corrupted with our Stain noise model to avoid convergence towards an identity operator. This new approach performs significantly better than traditional autoencoders to detect real defects on texture images of the MVTec AD dataset.

Furthermore, we also provided a fair comparison between the residual- and uncertainty-based detection strategies relying on our AESc + Stain model. Unlike the reconstruction residual, the uncertainty indicator is independent of the contrast between the defect and its surroundings, which is particularly relevant for low contrast defects localization. However, in comparison to the residual-based detection strategy, the uncertainty-based approach increases the false positive rate in the clean structures.

References

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," *Cvpr 2019*, pp. 9592–9600, 2019.

[2] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2013.12.026

[3] P. Seebock, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunovic, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT," *IEEE Transactions on Medical Imaging*, 2019.

[4] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 5575–5585, 2017.

[5] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," *arXiv preprint arXiv:1511.08861*, 2015.

[6] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, pp. 372–380, 2019.

[7] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially Learned One-Class Classifier for Novelty Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, 2018.

[8] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, pp. 146–147, mar 2017. [Online]. Available: http://arxiv.org/abs/1703.05921

[9] 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, no. January, pp. 30–44, 2019.

[10] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection," *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019. [Online]. Available: http://arxiv.org/abs/1901.08954

[11] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11383 LNCS, pp. 161–169, 2019.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[14] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4277–4281.

[15] C. Huang, J. Cao, F. Ye, M. Li, Y. Zhang, and C. Lu, "Inverse-Transform AutoEncoder for Anomaly Detection," *arXiv preprint arXiv:1911.10676*, 2019. [Online]. Available: http://arxiv.org/abs/1911.10676

[16] V. Chandola, A. Banerjee, and V. Kumar, "Survey of Anomaly Detection," *ACM Computing Survey*, no. September, pp. 1–72, 2009. [Online]. Available: http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf

[17] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by CNN-based self-similarity," *Sensors (Switzerland)*, vol. 18, no. 1, 2018.

[18] B. Staar, M. Lütjen, and M. Freitag, "Anomaly detection with convolutional neural networks for industrial surface inspection," *Procedia CIRP*, vol. 79, no. January 2019, pp. 484–489, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212827119302409

[19] C. Zhou and R. C. Paffenroth, "Anomaly Detection with Robust Deep Autoencoders," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, 2017.

[20] D. Dehaene, O. Frigo, S. Combrexelle, and P. Eline, "Iterative energy-based projection on a normal data manifold for anomaly localization," pp. 1–17, 2020. [Online]. Available: http://arxiv.org/abs/2002.03734

[21] M. Haselmann, D. P. Gruber, and P. Tabatabai, "Anomaly Detection Using Deep Learning Based Image Completion," *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pp. 1237–1242, 2019.

[22] A. Munawar and C. Creusot, "Structural inpainting of road patches for anomaly detection," *Proceedings of the 14th IAPR International Conference on Machine Vision Applications, MVA 2015*, pp. 41–44, 2015.

[23] S. Mei, Y. Wang, and G. Wen, "Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model," *Sensors (Switzerland)*, vol. 18, no. 4, pp. 1–18, 2018.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[25] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 001–13 008, 2020.

[26] R. Fong and A. Vedaldi, "Occlusions for effective data augmentation in image classification," *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pp. 4158–4166, 2019.

[27] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11215 LNCS, pp. 89–105, 2018.