Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 107 (2022) 246–251

www.elsevier.com/locate/procedia

55th CIRP Conference on Manufacturing Systems

Anomaly detection for industrial surface inspection: application in maintenance of aircraft components

Falko Kähler*[a], Ole Schmedemann[a], Thorsten Schüppstuhl[a]

[a]*Hamburg University of Technology, TUHH, Institute of Aircraft Production Technology, Denickestr. 17, 21073 Hamburg, Germany*

* Corresponding author. Tel.: +49-40-42878-3479 ; fax: +49-40-42731-4551. *E-mail address:* f.kaehler@tuhh.de

Abstract

Surface defects on aircraft landing gear components represent a deviation from a normal state. Visual inspection is a safety-critical, but recurring task with automation aspiration through machine vision. Various rare occurring faults make acquisition of appropriate training data cumbersome, which represents a major challenge for artificial intelligence-based optical inspection. In this paper, we apply an anomaly detection approach based on a convolutional autoencoder for defect detection during inspection to encounter the challenge of lacking and biased training data. Results indicated the potential of this approach to assist the inspector, but improvements are required for a deployment.

Keywords: optical inspection; anomaly detection; surface defects; machine vision

1. Introduction

Visual inspection is a repetitive task in product lifecycles. During maintenance of aircraft landing gears, the complete surface of components must be inspected reliably for surface defects to ensure flight safety. Small and hard to identify defects like pitting corrosion require skilled and experienced workers, which are prone to error and human factors such as monotony or unsteady concentration [9, 16]. Additional drawbacks of visual inspection such as hardly accessible component areas motivate the use of imaging sensors and automatic evaluation (machine vision) using artificial intelligence (AI), which is successfully deployed in a variety of applications.

AI-bases machine vision relies on sufficient training data that encompasses all possible types, shapes and locations of defects. Rarely occurring defect types and variations as well as missing or incomplete documentation make the creation of comprehensive datasets tedious, time-consuming and costly, representing a major challenge for machine vision deployment. Approaches to generate synthetic training data yield potential to deal with this challenge [4, 10, 17], but require expert knowledge and have a risk of unintentional domain gap. However, defect-free samples are easy to acquire in large quantities. This motivates to train AI-based approaches only on one class, the normal/defect-free samples. Ideally, the AI-models learn the structure of the normal class and subsequently identify anomalous samples when they sufficiently deviate from the learned normal state.

The goal of this work is to develop and implement an AI-based approach for detecting corrosion surface defects on landing gear components in image data based on one-class classification. After a review of related work and relevant aspects, an anomaly detection approach is selected and applied. Later on, results are discussed and the achieved performance is evaluated regarding deployment on the given use case.

Nomenclature

A	image A (input)
a	threshold parameter
B	image B (reconstruction)
i	pixel coordinate
j	pixel coordinate
n_F	number of filters
T	threshold
μ	mean value
σ	standard deviation

2212-8271
10.1016/j.procir.2022.05.197

2. Related work

2.1. Automated optical inspection

Application of automated optical inspection is often motivated by the demand for increased productivity and reduction of errors and costs. A broad field of application is the detection of defects, which aims to identify the specific class and location of a defect [11]. Most applications use an imaging system (image sensor and lighting) to capture the surface of the object with adjacent software for image evaluation. The software consist of an inspection algorithm to extract features of the image and classify it into non-defect or defect [9]. In recent years, inspection algorithms are more and more based on artificial intelligence, which has improved the performance of various computer vision tasks. Automated optical inspection in order to detect surface defects is applied in manufacturing quality control, such as of metal [11, 15], ceramics or textiles [8], optical elements [16] or electronics [9].

Another wide field is inspection during maintenance (as in our use case) with many researches been conducted up to now. For instance, AI-based inspection is more and more entering in aircraft maintenance. Inspecting the fuselage for corrosion is a vital task, where Brandoli et al. [5] applied a image-based deep learning method for corrosion identification and achieved promising results and high performance. The authors encountered shortage of defective images by employing transfer learning, but stated their method is expected to improve with more data. Taheritanjani et al. [19] applied supervised and unsupervised AI-methods on real image data of aircraft engines fasteners and achieved an accuracy and recall of 0.99 using a Resnet101-based supervised method, while unsupervised methods like support vector machines or autoencoders achieved significantly lower performance. The authors stated the main drawbacks of supervised methods are tedious data collection as well as the lacking generalizability when introducing unknown defects. Other researches used AI-based approaches on endoscopic images for defect detection in of aircraft engines [18, 21]. Shen et al. [18] for instance successfully implemented supervised learning for detecting cracks and burns, but the amount of available training data remains a bottleneck.

2.2. Anomaly detection

Classification between a defect-free and defective while training only on defect-free/normal data instances is often referred as one-class classification or anomaly detection problem [6, 7]. Various semi-/unsupervised approaches have been developed to encounter the data shortage issue. In recent years, approaches mostly based on a Generative Adversarial Network (GAN) or Convolutional Autoencoder (CAE) have been developed to detect defects in image data. However, compared to traditional classification approaches, the specific type of defect cannot be determined. An et al. [3] proposed a Variational Autoencoder (VAE), which differs from conventional autoencoders that it delivers a reconstruction probability instead of a reconstruction error, which does not require a specific threshold for classification. According to their study, the proposed method outperformed autoencoder or principal component analysis. Tsai et al. [20] developed a convolutional autoencoder (CAE) approach for defect detection and applied it successfully on a variety of material surfaces. Their introduced CAE with regularizations outperformed conventional CAE as well as VAE. Lehr et al. [14] compared a CAE with pretrained and fine-tuned convolutional neural networks (ResNet-18 based) on their own data as well as on MVTec dataset. They observed the CAE performed better detecting defective images than defect-free on their own created dataset. However, compared to supervised methods, the CAE achieved lower accuracy than supervised methods. GAN-based approaches have received increasing attention from researchers in recent years [1, 2, 12]. For instance, Lai et al. [12] used a pretrained GAN to generate defect-free images based on their training data. As the proposed GAN failed to generate defective samples, they were able to identify defects in textured images effectively.

3. Use case analysis

According to [6, 7], different aspects of anomaly detection (see fig. 1) have to be discussed to select appropriate anomaly detection methods. Regarding input data, it can be distinguished

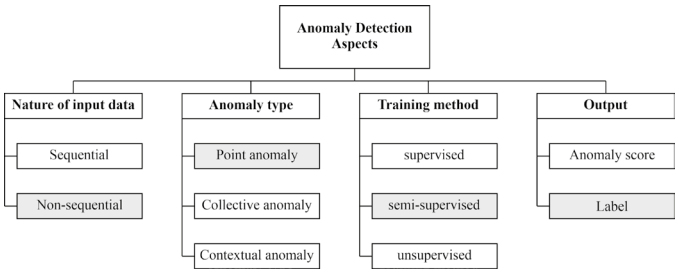

Fig. 1. Aspects of anomaly detection.

between sequential data (e.g. video, speech, text) and non-sequential data (e.g. images). This work focuses on imaging sensors to capture surface defects, so non-sequential image data is used. Due to lacking defect data, only one class data of defect-free images is available for training. Therefore, only semi-supervised methods are considered, which only train on one class (one-class classification). Next, anomalies can be categorized into three types:

- Point anomalies, where a data instance significantly deviates from the other instances
- Contextual anomalies, where a data instance is considered anomalous in a specific context
- Collective anomalies, where single data instances appear to be normal, but anomalous in a group

Pitting corrosion defects on landing gears can be considered point anomalies since they do not have a specific context between each other as their occurrence is random. They are individually considered anomalous. Output of deep anomaly detec-

tion methods can either be an anomaly score or a class label. An anomaly score quantifies the outlierness of a data instance. The score can be ranked and a domain-specific threshold (decision score) determined by an expert can be applied to identify anomalies [6]. As our goal is to detect surface defects, we aim for a label output. Anomaly scores, however, may yield useful information about the defectiveness.

4. Approach

After reviewing aspects of anomaly detection in section 3, a Convolutional Autoencoder (CAE) approach has been selected. An autoencoder consists of an encoder and decoder. The encoder compresses the input to a lower-dimensional space. This compressed representation is passed to the decoder, which reconstructs it back to the input dimension. For CAE, the encoder consists of a sequence of convolution and downsampling layers to compress an image, while the decoder involves a series of deconvolution and upsampling layers for reconstruction [20]. Compared to VAE- or GAN-based approaches, CAE are considered relatively straightforward to train.

As a generative method, the CAE delivers an image output which can be compared with the input. The similarity or reconstruction error between input and output characterizes the autoencoder's performance and is considered as anomaly score. Since the CAE will be trained only on reconstructing defect-free images, a high similarity between input and reconstructed output and therefore a low reconstruction error is expected. For defective images, a higher reconstruction error is estimated, which deviates significantly from normal instances. We propose a threshold method based on the reconstruction error to identify normal and anomalous instances. Different metrics for calculating the similarity of input and reconstruction, namely mean squared error (MSE), structural similarity index (SSIM) and signal-to-reconstruction-error-ratio (SRE) [13], are considered.

5. Implementation

5.1. Data acquisition

Real images of a landing gear component surface have been acquired for training and testing. Figure 2 depicts the acquisition setup which consists of a grayscale camera focusing perpendicular on the component surface. The component itself can be rotated while the camera is slideable parallel along the component rotation axis. Due to the varying outer component contour, the camera was manually focused. A LED ring light ensured adequate lighting conditions. In total, 600 non-defect and 300 defect images were taken. The images were resized to 144x144 pixels and normalized. Dataset samples[1] are shown in fig 3. Corrosion is visible as dark areas, which clearly separates from the (rather noisy) metal texture.

[1] The dataset can be requested from the corresponding author.

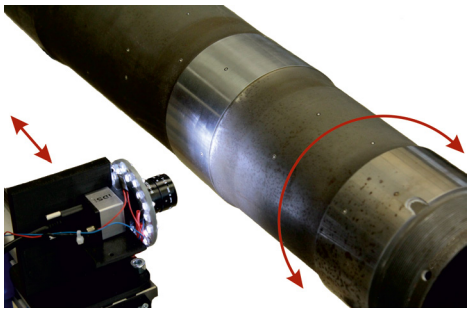

Fig. 2. Setup for image acquisition.

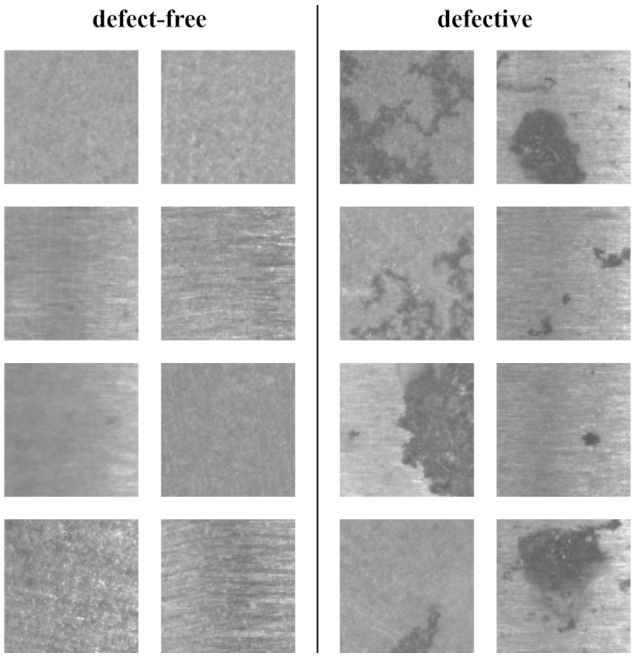

Fig. 3. Dataset samples.

5.2. Autoencoder architecture

The CAE was implemented using python and keras tensorflow. The encoder consists of the input and a convolution layer to compress the input image. The decoder applies a transposed convolution and a convolution layer to reconstruct the input from the compressed representation. The width of the autoencoder is varied by modifying the number of filters (n_F) of the convolution/convolution transpose layers between 16, 32 and 64. Adam (short for Adaptive Moment Estimation) optimizer was applied. Details of the architecture can be found in table 1.

Table 1. Detailed parameters of the CAE.

	Layer	Filters	Padding	Activation
Encoder	Input			
	Convolution	n_Fx(3,3)	same	relu
Decoder	Convolution transpose	n_Fx(3,3)	same	relu
	Convolution	1x(3,3)	same	sigmoid

5.3. Training

The CAE models were trained from ground up using the data elaborated in section 5.1. 500 random defect-free samples were picked for training and the remaining 100 defect-free and all 300 defective samples were used for subsequent testing. During training, a 85/15 training/validation split was applied. The training data is shuffled after each epoch. Training parameters were chosen on best practices among literature. A constant batch size of 16 and a learning rate of 0.001 were chosen. All models were trained for 25 epochs and saved after each epoch for evaluation.

6. Results

6.1. Selection of autoencoder architecture and metric

The CAE models have been trained and the performance (balanced accuracy, precision and recall) on the test samples was calculated. The similarity measures were compared to determine the best metric for normal/defective classification. As the majority of defect-free training samples would be classified correctly using the mean reconstruction loss, it is used as initial threshold T for classification of test samples, as a significantly differing loss is expected for defective samples. Hereby, for MSE losses greater than the threshold are considered anomalous (less error equals greater similarity), while for SSIM and SRE losses lower than the threshold are considered anomalous (higher ratio equals greater similarity).

Figure 4 shows the performance of each training configuration over the number of epochs. Despite the shallow CAE architecture, promising, but not yet sufficient results are achieved at this stage. Best performance is achieved for $n_F = 64$ after one epoch and $n_F = 16$ after 3 epochs using SRE metric. For MSE, individual performance values reached similar levels. It can be noticed SSIM lacks behind and is not considered further in this work. Table 2 shows the respective confusion matrices for $n_F = 64$ and $n_F = 16$.

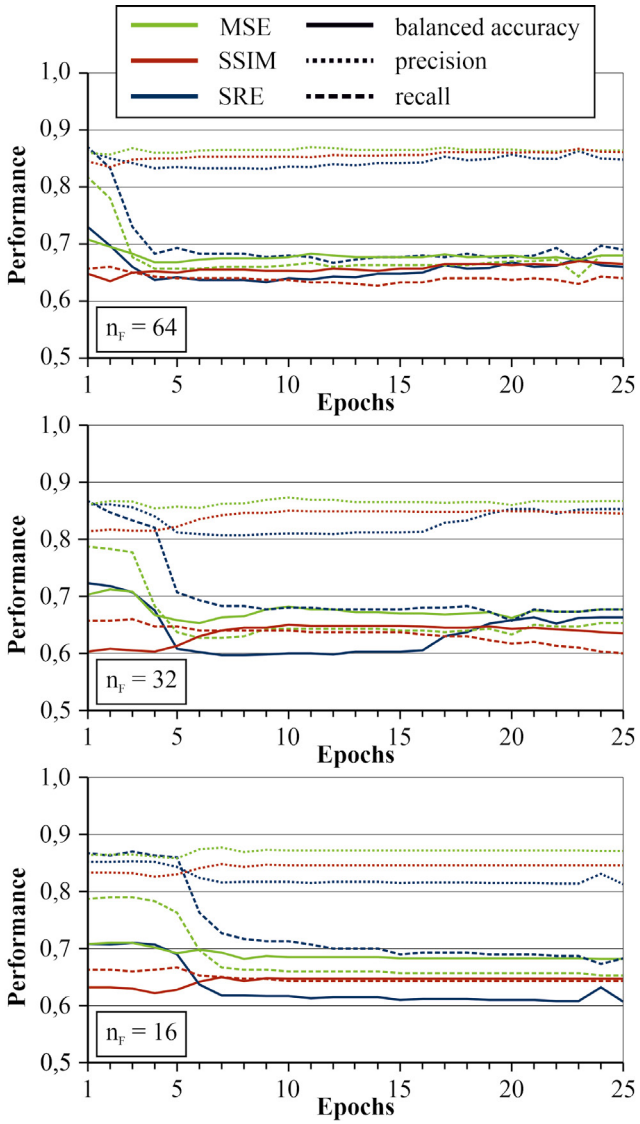

Fig. 4. Performance of MSE, SSIM and SRE metric with varying number of filters.

Table 2. Confusion matrix for $n_F = 64$ after 1 epoch (top) and $n_F = 16$ after 3 epochs (bottom).

$n_F = 64$		Predicted label: defective	Predicted label: defect-free	Total
True label	defective	261	39	300
	defect-free	41	59	100
	Total	302	98	400

$n_F = 16$		Predicted label: defective	Predicted label: defect-free	Total
True label	defective	261	39	300
	defect-free	45	55	100
	Total	306	94	400

Roughly 50% of defect-free samples were falsely classified, indicating the initial threshold is not suitable for evaluating defect-free samples. In contrast, defective samples were classified significantly better. Since all defects must be reliably detected when inspecting safety-critical components, a high recall is desired and as small number of false positive detections may be accepted as a trade-off. Due to overall highest recall rates, SRE metric and $n_F = 64$ is selected for further investigations and optimizations such as threshold adjustment.

6.2. Threshold adjustment

The decision threshold is important for the model performance. In section 6.1, the mean training reconstruction error was used as initial threshold to determine a suitable similarity metric. However, tuning the threshold may reinforce the desired behavior for the specific use case. We define the threshold as $T = \mu + a \cdot \sigma$, where the initial mean reconstruction error μ is adjusted by the product of parameter a and the standard deviation σ of the training reconstruction errors. The wide CAE-model ($n_F = 64$) after one epoch is used for further evaluation. Parameter a is varied between -2 and 2 and the performance is evaluated. As depicted in figure 5, decent tradeoffs of all three

performances are achieved in range $0 \leq a \leq 0.5$. One can also

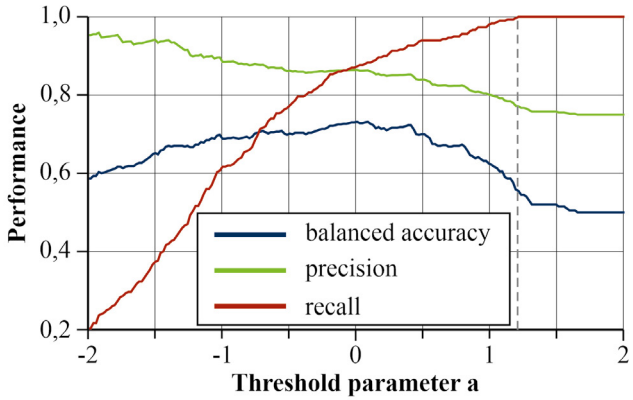

Fig. 5. Threshold adjustment.

notice the contrary behavior of precision and recall as a high recall leads to more defect classifications, but also to more incorrect defect classifications, decreasing the overall precision. As the slopes of precision and recall indicate, a high increase of recall can be achieved while sparsely decreasing precision. Since inspection is a safety-critical task, all defects must be detected, so a high recall rate is desired. When adjusting the threshold to $a = 1.21$, a maximum recall rate of 1 is achieved, but balanced accuracy and precision decrease to 0.56 and 0.77 respectively, according to the corresponding confusion matrix in table 3. The performance decreased drastically, with only 12

Table 3. Confusion matrix. ($n_F = 64$, epoch = 1, $a = 1.21$).

		Predicted label		
		defective	defect-free	Total
True label	defective	300	0	300
	defect-free	88	12	100
	Total	388	12	400

of 100 defect-free samples classified correctly, resulting only in a 12 % reduction of images to be evaluated manually. The false alarm rate would be extremely high in real deployment, as defect-free samples are more frequent than defective. In this state, it is evident a threshold adjustment does not lead to satisfactory results and the approach does not yield sufficient benefit for the inspector yet.

6.3. Failure investigation

In order to identify causes for the low performance, we investigated the reconstructions of the CAE. Figure 6 shows data samples from the test set, both normal and defective. As the images indicate, the CAE successfully reconstructs input from both normal and defective samples and visual differences are not as evident expected. The reconstructions appear blurry probably causing a distorted similarity measurement. Brightness differences can be noticed between input and reconstruction for both normal and defective samples. We conclude the CAE (with the elaborated architecture) has not learned properly on the given data. The training data was analyzed for sus-

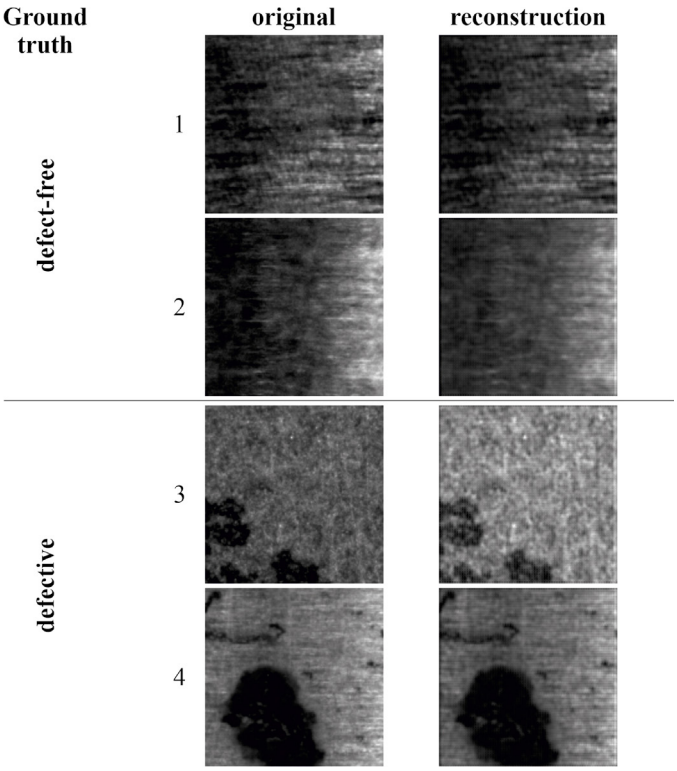

Fig. 6. Image reconstructions.

pected issues and bias. Some blurry and probably misleading samples were noticed. For instance, original image 1 in figure 6 shows ambiguous dark spots, which were not considered defective during data acquisition. We investigated falsely classified defect-free images (fig. 7), which reinforced this suspicion. As the images were normalized during data preparation, dark spots (see red boxes in fig. 7) become more present. These spots have similarities with the actual corrosion defects and might mislead the CAE preventing it from learning relevant class features, resulting in a hazy demarcation line between normal and defective.

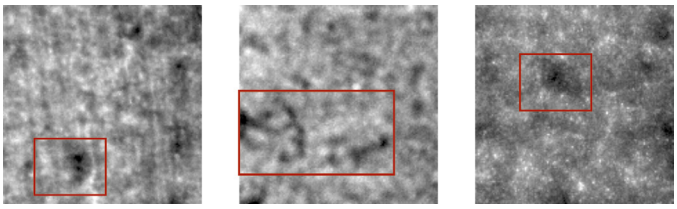

Fig. 7. Falsely classified defect-free samples.

6.4. Discussion

As previously elaborated, the CAE reconstructs defect-free as well as defective images successfully, which influences the performance negatively. On one hand, threshold adjustment according to section 6.2 can increase the desired behavior of the CAE to detect all defects, but on the other hand will drastically increase the false positive classifications and, in contrast

to previous works elaborated in section 2.2, does not lead to a satisfactory performance for autonomous deployment. The approach can be considered as starting point for machine-vision based inspection and yield potential to reduce the number of images to be evaluated manually and the time spent for visual inspection when the false alarm rate is on an acceptable low level. To achieve this, the training data could be enhanced and expanded. This is on one hand accomplished by careful data acquisition to describe the normal, defect-free state as precisely and unambiguously as possible to separate anomalies clearly. Deeper information content such as colored images may support this separation. On the other hand, during deployment the data itself can could be expanded by new data annotated by the still-needed inspector in aircraft landing gear maintenance. This enables an further algorithm training and improvements, such as inclusion of new or variable over time normal states.

7. Conclusion

This work investigates a method to detect surface defects in image data of an aircraft landing gear component. To encounter the shortage of defective samples, we pursued an one class anomaly detection approach based on a convolutional autoencoder (CAE). The CAE is expected to successfully reconstruct normal/defect-free images, but fail on anomalous/defective samples. The similarity between input and reconstructed output is calculated and compared with a threshold to identify defective samples. Results showed the implemented CAE reconstructs normal as well as defective inputs successfully, which affects the performance negatively. Several metrics for evaluating the similarity and reconstruction error have been investigated, with signal-to-reconstruction-error (SRE) proving to be the most effective metric to differentiate between normal/defective. However, due to successful reconstruction of defective samples, a clear threshold could not be determined and a satisfactory performance is not achieved. The approach can be considered as starting point for machine vision based surface inspection. An adjusted CAE architecture, higher quality training data as well as gathered data during deployment and probably additional information such as color might increase the performance and yield potential to reduce the number images to be evaluated manually.

Acknowledgements

Research was funded by the German Federal Ministry for Economics and Climate Action under the Program LuFo V-3.

References

[1] Akcay, S., Atapour-Abarghouei, A., Breckon, T.P., 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training .

[2] Akçay, S., Atapour-Abarghouei, A., Breckon, T.P., 2019. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection.

[3] An, J., Cho, S., 2015. Variational autoencoder based anomaly detection.

[4] Bath, L., Schmedemann, O., Schüppstuhl, T., 2021. Automatisierung in der industriellen Endoskopie/development of new means regarding sensor positioning and measurement data evaluation – automation of industrial endoscopy. wt Werkstattstechnik online 111, 644–649. doi:10.37544/1436-4980-2021-09-70.

[5] Brandoli, B., de Geus, A.R., Souza, J.R., Spadon, G., Soares, A., Rodrigues, J.F., Komorowski, J., Matwin, S., 2021. Aircraft fuselage corrosion detection using artificial intelligence. Sensors (Basel, Switzerland) 21. doi:10.3390/s21124026.

[6] Chalapathy, R., Chawla, S., 2019. Deep learning for anomaly detection: A survey. URL: http://arxiv.org/pdf/1901.03407v2.

[7] Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection. ACM Computing Surveys 41, 1–58. doi:10.1145/1541880.1541882.

[8] Czimmermann, T., Ciuti, G., Milazzo, M., Chiurazzi, M., Roccella, S., Oddo, C.M., Dario, P., 2020. Visual-based defect detection and classification approaches for industrial applications-a survey. Sensors (Basel, Switzerland) 20. doi:10.3390/s20051459.

[9] Ebayyeh, A.A.R.M.A., Mousavi, A., 2020. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. IEEE Access 8, 183192–183271. doi:10.1109/ACCESS.2020.3029127.

[10] Gutierrez, P., Luschkova, M., Cordier, A., Shukor, M., Schappert, M., Dahmen, T., 2021. Synthetic training data generation for deep learning based quality inspection doi:10.1117/12.2586824.

[11] He, Y., Song, K., Meng, Q., Yan, Y., 2020. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Transactions on Instrumentation and Measurement 69, 1493–1504. doi:10.1109/TIM.2019.2915404.

[12] Lai, Y.T., Hu, J.S., Tsai, Y.H., Chiu, W.Y., 09.07.2018 - 12.07.2018. Industrial anomaly detection and one-class classification using generative adversarial networks, in: 2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), IEEE. pp. 1444–1449. doi:10.1109/AIM.2018.8452228.

[13] Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K., 2018. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. ISPRS Journal of Photogrammetry and Remote Sensing 146, 305–319. doi:10.1016/j.isprsjprs.2018.09.018.

[14] Lehr, J., Sargsyan, A., Pape, M., Philipps, J., Krüger, J., 08.09.2020 - 11.09.2020. Automated optical inspection using anomaly detection and unsupervised defect clustering, in: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE. pp. 1235–1238. doi:10.1109/ETFA46521.2020.9212172.

[15] Luo, Q., Fang, X., Liu, L., Yang, C., Sun, Y., 2020. Automated visual defect detection for flat steel surface: A survey. IEEE Transactions on Instrumentation and Measurement 69, 626–644. doi:10.1109/TIM.2019.2963555.

[16] Schöch, A., Perez, P., Linz-Dittrich, S., Bach, C., Ziolek, C., 2017. Automated surface inspection of small customer-specific optical elements. tm - Technisches Messen 84, 502–511. doi:10.1515/teme-2017-0012.

[17] Schoepflin, D., Holst, D., Gomse, M., Schüppstuhl, T., 2021. Synthetic training data generation for visual object identification on load carriers, pp. 1257–1262. doi:10.1016/j.procir.2021.11.211.

[18] Shen, Z., Wan, X., Ye, F., Guan, X., Liu, S., . Deep learning based framework for automatic damage detection in aircraft engine borescope inspection , 1005–1010doi:10.1109/ICCNC.2019.8685593.

[19] Taheritanjani, S., Schoenfeld, R., Bruegge, B., 2019. Automatic damage detection of fasteners in overhaul processes, in: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), IEEE. pp. 1289–1295. doi:10.1109/COASE.2019.8843049.

[20] Tsai, D.M., Jen, P.H., 2021. Autoencoder-based anomaly detection for surface defect inspection. Advanced Engineering Informatics 48, 101272. doi:10.1016/j.aei.2021.101272.

[21] Wong, C.Y., Seshadri, P., Parks, G.T., 2021. Automatic borescope damage assessments for gas turbine blades via deep learning 142, 1097. doi:10.2514/6.2021-1488.