Project report

# Automatic Data Analysis within Non-Destructive Evaluation (ADA-NDE)

Project report for ÅForsk funded project number 19-546

*Erik Lindgren* *Christopher Zach*

2023 October

UNIVERSITY WEST

CHALMERS
UNIVERSITY OF TECHNOLOGY

**Abstract**

This is the project report for the ÅForsk funded project Automatic Data Analysis within Non-Destructive Evaluation (ADA-NDE). The project has been conducted in collaboration between University West and Chalmers University of Technology.

Recently studies have shown amazing results on utilizing Deep Learning-based machine learning models for image analysis, both in a wide applications domain as well as in non-destructive evaluation (NDE. However, most of the solutions are overconfident when subjected to test data highly dissimilar to the training data, so called out-of-distribution (OOD) data. We claim that it is important that computer algorithms and models for data interpretation in NDE of high value critical products react sensible also to unexpected rare OOD input.

I this project we have explored Deep Learning-based approaches to OOD detection, as an estimate of the level of a confidence in the results with respect to OOD data. We have focused on industrial X-ray inspection but also explored the results on the industrial visual inspection application. Convolutional neural networks were trained to model data image distribution of the training data, both for the OOD detection application but also for the generative application (generate new samples).

The physics of the X-ray image formation was taken into special consideration when deriving loss functions and augmenting the training data; for example considering challenges with spatial correlated X-ray quantum noise as well as the basic features of X-ray transmission imaging such as the signals representing the sums of attenuation along the X-ray paths.

The models were trained similar to Denoising Auto-Encoders, but instead of with Gaussian noise added, rather with highly structural noise (perturbations) added, representing anything outside of the training dataset. We call the set of models perturbed Auto-Encoders, somewhere in between unsupervised learning and supervised learning, with an intrinsically built-in sensible reaction to OOD data. In this project we successfully demonstrated that on specific example applications, within X-ray inspection and visual inspection, such perturbed AEs perform similar to conventional supervisedly trained DL models with test data similar to the training data and exceed in performance over those when subjected to OOD test data.

In addition, we have shown that it is possible to successfully train on hybrid datasets with real and synthetic data combined together.

Keywords: NDE; NDT; Machine Learning; Deep Learning;

# Contents

# 1 Introduction

Non-destructive evaluation (NDE) is a form of quality control where the evaluated objects are not destroyed by the control itself. The goal is to detect and characterize material deviations that have a negative effect on the properties of the object. NDE is used during manufacturing and later in the product life. NDE is especially extensively utilized within quality critical industries (e.g. nuclear, oil and gas, and aerospace) where unexpected product failures will lead to serious loss of human lives.

Within quality critical high value industries the analysis of the NDE measurement data and the subsequent interpretation to turn it into material quality decisions is often done manually by trained operators. Clearly, there is a strong interest within these industry segments to decrease this human effort in data analysis in order to be more efficient and more consistent in the interpretations. The paradigm shifts with Industry 4.0, further puts requirements on the NDE community to also move into the fourth industrial revolution [1], with high degrees of automation. Other trends such as the industrialization of metal additive manufacturing, 3D printed metal parts, and its inherent connection to a digital work flow at manufacturing, also calls for increasing the automatization in the NDE steps.

In this project we have explored semi-automatic approaches for data analysis rather than fully automatic ones. Rather than fully removing the Human, we have the vision that Human and Artificial Intelligence (AI) will work together in close collaboration when it comes to data analysis, or interpretation, within critical applications of NDE. Humans prefer a certain amount of variation in the data (and tasks) to avoid loss of focus and attention, and humans are good at spotting new and unusual (visual) elements not seen before; while getting tired and error prone when there is too little variation. The opposite is true for machine learning-based (ML) AI, which today even excel over humans in many image interpretation tasks, at least as long as the image to be tested is similar to the data it was trained on. However, when the data is dissimilar to the training data, most of the ML-based algorithms will be overconfident in their results, with potentially unsafe outcomes. We therefore propose that a safe and effective collaboration between Human and AI, in this application area, is relying on a sound confidence estimation of the AI, especially with respect to new unexpected data far from the training data. With such a measure ML-based AI can take a large amount of the workload off the human, since most of the data will be similar to the training data; but request assistance from the human with cases that are dissimilar to the training data, for which it does not know enough to make a safe interpretation by itself.

In this work we have explored how an accurate and conservative estimation of confidences (how certain the algorithm is), with respect to new unexpected data far from the training data, can be achieved. Such a confidence estimation can then be used to judge when a human operator should get involved in further decision making and when the results from the AI are sufficiently reliable and safe to use directly. In this project report, we will summarize the most important results which have already been published in the project. For details, we refer to the publications produced in the project [2, 3, 4, 5].

## 1.1 Project plan

Below the work packages and milestones from the project funding application are given.

**Work package 1: X-ray inspection, data and interpretation model**
Model for analysis results confidences estimate and a model for the data (synthetic radiographs), both closely connected. Both are input to work package 2.
Deliverables: Data and interpretation model, experimental and synthetic data.

**Work package 2: Machine learning algorithms**
Explore active machine learning algorithms, derive and implement proof-of-concepts and compare their capability. The focus is on algorithms that estimate reliable confidences for out-of-distribution inputs.
Deliverables: Journal paper, conference paper, master thesis.

**Work package 3: Generalization to other NDE methods and applications**
Generalize WP1 and WP2 results to other methods, e.g. laser ultrasound, thermography, visual inspection.
Deliverables: Conference paper

**Milestones** (year and number indicate what half of the year):

- 2020, 1: Master student thesis work; conference paper
- 2021, 1: Conference paper
- 2021, 2: Journal paper (main results)
- 2022, 1: Public report and conference paper (generalization)

All publication milestones were delivered; however, the project did not succeed to deliver a master thesis.

# 2 Background

Performing industrial inspection can be divided into three steps: planning, data sampling, and data interpretation. The automatization of the interpretation step is one of the main bottlenecks for quality-critical industries. By interpretation, we mean for example in the case of X-ray inspection, the operation of transforming (mainly) the X-ray image data into information that can be utilized later for decisions, e.g., decisions regarding the material quality. We believe that understanding of the physical processes involved in the formation of the measurement data is important in this interpretation automatization context. Therefore, in this project a single NDE method was chosen to focus on for the proof-of-concept.

Most of the automatic industrial X-ray image interpretation algorithms in literature have some steps in common: pre-processing, segmentation, feature extraction, and classification. The preprocessing step typically includes image data calibration, noise removal, and contrast enhancement. In the segmentation step, the image is separated in different parts, most often the pixels belonging to potential material defects or indications. Over the year many different segmentation algorithms have been explored, see for example [6] for a comprehensive study on different segmentation methods for metal fusion-weld defect segmentation.

In the feature extraction and classification step, each segmented region is described by a feature vector which is then utilized to classify the region. This art of feature engineering has traditionally been done manually; however, much of the recent machine learning breakthroughs are based on Deep Learning, where the features are instead learned automatically in training stage. In this project we have focused on such DL-based solutions. An introduction to DL is outside the scope of this report, see for example [7] for an introduction.

The case of manually derived features, followed by a supervised classifier, has been extensively studied for the application of X-ray inspection of metal fusion-welds and aluminium castings; e.g. fuzzy logic [8], contrast-variance features [9], Support Vector Machines [10], Artificial Neural Networks [11, 12] (ANN), and Random forest classifiers [13]. Two conclusions can be drawn from earlier studies: first, a high accuracy in the results delivered by the models is possible when tested on data similar to the training data, and second, there is tendency for features related to local variation in the images, to perform well.

For the DL approach, the number of studies on exploring it for the analysis of industrial X-ray images is growing fast. Some example studies are: Weld defect classification [14, 15, 16], weld defect segmentation [17, 18], image patch level classification of aluminium castings [19], and X-ray computed tomography (XCT) segmentation in [20]. Also on industrially relevant visual inspection data, DL models have shown very promising results [21, 22, 23, 24]. Overall they show very promising performance, indicating very high detection rates at low false positives rate; as long at the test data is similar to the training data.

In additional to these heavily explored approaches consisting of the four steps described above, there are also studies on a more statistical hypothesis testing or residual analysis framework, hence closer to an anomaly detection approach. Where deviations from a training set containing X-ray images of accepted material quality are identified as anomalies or potential material imperfections or defects. Early studies can be found in [25, 26, 27, 28]. Recently, DL-based approaches have also started to attract attention within this field of anomaly detection. As in [29] on industrial X-ray CT data and in [30] on industrial X-ray inspection of die casts; or, as in our first publication in this project [2]. Further, DL-based anomaly detection approaches are gaining quite a lot of attention in a broader industrial image analysis domain [31, 32]. These DL models are essentially trained unsupervisedly (only with the accepted image distributions, one single class of data) in contrast to the supervisedly trained DL models which are trained to explicitly discern between different classes in the training data. The resource intense problem of manually labeling data (e.g. creating segmentations, or image classification) thus can be minimized or more or less excluded; apart from intrinsically addressing the challenge with OOD data at test time.

In summary, current state-of-the-art studies are DL-based approaches which show promising results with high accuracy on test data similar to training data. However, very few of the studies have explicitly addressed how the algorithms react when subjected to unexpected new input data far from the training distribution, so called out-of-distribution (OOD) data. We claim that, for industrial NDE applications on critical products, a confidence estimation with respect to such OOD data is important and should be addressed. In this project we have addressed such confidence estimation with respect to OOD data, as well as taking the OOD challenge into consideration when training and designing the proposed DL models.

# 3 Project results

The core question in this project has been about how to achieve a safe estimate of the confidence in the results, at inference time, with respect to unexpected OOD data. This question was further refined into how to derive an OOD data detection model, which is essentially a one-class classifier.

The core idea we explored was to use Deep Learning-based methods to model the accepted input distribution, much like a filter accepting input similar to the training data and rejecting data dissimilar to the training data. Any rejected data would then indicate the presence of OOD data. Closely related to this, or another view on it, is the ability of a model to generate data similar to a training dataset. Therefore we have also explored generative models, where new samples are drawn from the modeled training data distribution.

As planned in the project application, the specific challenges (and possibilities) with a selected class of NDE methods was focused on, namely X-ray-based transmission imaging methods.

All of the results in the project have been published in four scientific publications:

Publication A: Autoencoder-Based Anomaly Detection in Industrial X-ray Images [2],

Publication B: Analysis of industrial X-ray computed tomography data with deep neural networks [3],

Publication C: Industrial X-ray Image Analysis with Deep Neural Networks Robust to Unexpected Input Data [4],

Publication D: Deep-learning-based out-of distribution data detection in visual inspection images [5].

What follows is a summary of the methods, models, datasets, results, and conclusions of these publications. For details we refer to the specific publication.

## 3.1 OOD detection (publications A-D)

Our approach for OOD detection was to model the accepted input distribution at a high precision, and to detect OOD data as inputs that were poorly modelled in the test image. The main method to model the accepted input data were based on auto-encoders (AEs), which are deep neural networks trained to reconstruct the accepted input, but in our case also to reject to reconstruct data outside of the accepted input distribution. At inference time, a large difference between input image and the reconstructed image (residual image) signals OOD data, or a localized anomaly. The residual image analysis was kept simple and related to the image formation physics for the X-ray images, essentially thresholding on local absolute value and standard deviation.

Publication A focused on the application of industrial 2D X-ray inspection of metal fusion welds welds. Compared to previous studies we added a localized reconstruction loss term while training the AE model. The AE model was also trained on both accepted as well as on systematic noise from not accepted images, in order to better handle not accepted input very similar to accepted input; i.e. welds of accepted quality contain image structures highly similar to welds with not accepted quality. We achieved a true positive rate at some $80 - 90\,\%$ at a false positive rate at around $4\,\%$, and at the same time correctly detected an OOD data example.

In Publication B the application was still X-ray-based image data, however 3D instead of 2D, with the application X-ray computed tomography (XCT) evaluation of additively manufactured metal. Specific XCT image related challenge with spatially correlated X-ray quantum noise in the images was addressed. In addition to Publication A we explored if X-ray noise augmentation during the AE-training could be used to constrain the AE generalization capabilities.

In Publication C we formalized the results from the two earlier publications. Essentially the same OOD detector architecture and models were utilized. However, an approach to utilizing a synthetic dataset for the systematic noise (perturbation dataset), derived partly from a completely different application domain (not X-ray images), was explored. The application was again 2D X-ray inspection of metal welds. We also derived and trained a conventional supervisedly trained DL model to explicitly illustrate its shortcomings with respect to OOD data at inference time. In addition, models of OOD data examples, with better control of its variation, was derived for evaluating the proposed OOD detector. We achieved greatly improved performance with true positive rates at around $90\,\%$ at false positive rates at around $0.1\,\%$ on samples similar to the training data and correctly detected some example OOD data.

In Publication D we generalized the results from the three first publications on X-ray-based image data to another NDE application, the visual inspection of metal surfaces. A publicly available dataset was utilized and our results were compared to other studies. Our proposed models perform worse than binary classifiers trained supervised on in-distribution test data. However, a performance gain compared to earlier studies on models trained unsupervised was indicated.

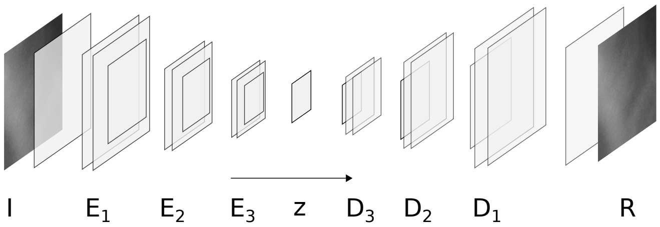

Figure 1: Illustration of the AE model with its encoder (E$_i$) and decoder (D$_i$) layers. Reused from [2] with permission from ASME.

### 3.1.1 Model architecture

The OOD detector is based on modelling the accepted input distribution with a reconstruction model and to detect OOD data as input failing to be reconstructed by the model. At inference time, the input test image is fed through the reconstruction model and its output (the reconstruction) is then subtracted from the input image to form a residual image. The residual image is then analyzed for large deviations with a simple physics-based model of the expected X-ray noise distribution. If the reconstruction errors are smaller than, or similar to the X-ray noise levels, then OOD data, or anomalies, can be detected at reasonable false positive rates.

At the heart of the proposed approach is the AE model, illustrated in Fig. 1. It is a deep convolutional neural network, consisting of an encoder part, which down-samples and compresses the input into latent space ($\mathbf{z}$), and a decoder part which reconstructs the input from the latent space representation. Between the down-sampling steps there is a varying number of convolution layers.

The AE models are kept small in size (some $50000 - 150000$ parameters), in order to make them intrinsically bad at generalizing to reconstructing input far from its accepted input distribution; i.e. to reject to reconstruct data outside of the training dataset. The models are similar in all publications, in Publication A three down-sampling stages are utilized, in B and C two down-sampling steps, and in C the number of convolution layers in series is higher. In Publication D we also explore the effect of increasing the dimension of latent space, leading to less compression.

### 3.1.2 Model training

The AE model is trained similar to a Denoising Auto-Encoder (DAE). A DAE is trained with the objective to recover low noise images from noisy images. Conventionally, they are trained by adding simple (e.g. Gaussian) noise to its input image, which then should be reconstructed without the added noise. In our case, we add highly structural noise (perturbations) to images with accepted intensity distributions, the noise representing anomalies or material imperfections which should not be reconstructed. To discriminate our model from the DAE:s we call it a perturbed AE ($\delta$AE) but essentially it is a DAE trained with structured noise.

We utilize a loss function (what to be minimized during the training) with multiple terms. The first term is given by the average square deviations of the reconstruction errors, the input image subtracted from the reconstructed image. However, such a loss tends to create reconstructions that are smooth and fail to reconstruct high frequency content in the images. Therefore, additional loss terms were explored; one term with a max-norm inspired loss, given by the maximum reconstruction errors, as well as several different kernel-like local loss terms.

The kernel-like terms were built by sliding a small kernel over all of the pixels in the reconstruction error image. Then taking the average or maximum over those slided kernels.

Inside the kernels, different operators would be applied; in Publication A, the average within the kernel, and in Publication B, the spread or range of intensity values within the kernel.

The training approach described can be classified as unsupervised training, i.e. the models are trained only on unlabelled data from one class; at least as long as the perturbation dataset is very general with weak prior assumptions. However, as the project proceeded we opted more and more towards classifying our training rather as supervised training instead. Since, we make use of labelled data in the training and the assumptions on the perturbation dataset were growing.

A large difference between conventional supervisedly trained binary classifier DL models (e.g. segmentation with a UNet-like model [33]) and our proposed $\delta$AE is that we explicitly force our model to be able to represent and reconstruct (in a sense understand) the whole accepted image intensity distribution while at the same time not reconstructing the perturbations; whereas the conventional supervisedly trained ones might be argued to be forced only to represent (in a sense understand) parts of the input. Therefore, the $\delta$AE approach has potentially a better intrinsic sensible reaction to OOD data at test time. This we demonstrated in publication C by comparing it to a supervisedly trained more conventional model.

### 3.1.3 Dataset and data augmentation

We decided to utilize publicly available datasets, instead of as planned, to derive our own datasets. The two main reasons were that we concluded that we would put too much resources on a) derive those datasets and b) implementing the algorithms of other studies and evaluate those on our new dataset, to be able to compare our results.

The public dataset GDXray [34], with digitized X-ray images of metal fusion-welds, was utilized in Publications A and C. On the downside it is limited in data size, and consists of digitized X-ray images rather than more modern digital X-ray images made digital directly with so called digital detector arrays. On the upside, it is a seriously derived popular dataset in the X-ray inspection community. In Fig. 2 an example weld image can be seen, and in Fig. 3 examples of X-ray image patches of welds with accepted as well as not accepted quality are shown.

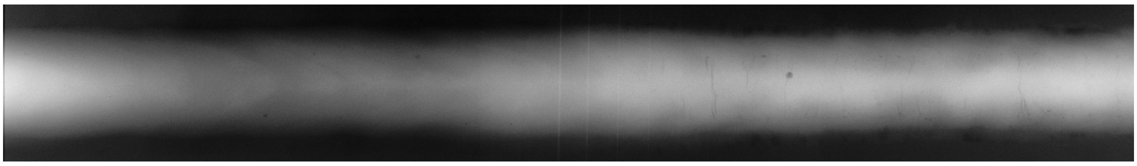

Figure 2: Example of an X-ray image cropped to the weld region. The white vertical line indications are from an image quality indicator, and both crack and pore indications can be seen. The X-ray image comes from the GDXray dataset [34]

The X-ray computed tomography dataset utilized in Publication B, is publicly available, and consists of 3D data volumes of CT evaluations of additively manufactured metal with material imperfections [35]. Two examples of 2D CT slices can be seen in Fig. 4.

In Publication D the publicly available Kolektor Surface Defects dataset[21], a visual inspection dataset, was utilized. It consists of visual images of metal surfaces, with and without crack defects present. Some examples can be seen in Fig. 5 and Fig. 6.

As indicated in the section on training of the AE models, perturbation datasets were utilized during training. In Publication A only real indications, extracted from their original weld radiograph and then inserted into the image patches with accepted weld material imaged, were utilized. Some examples can be seen in Fig. 7.

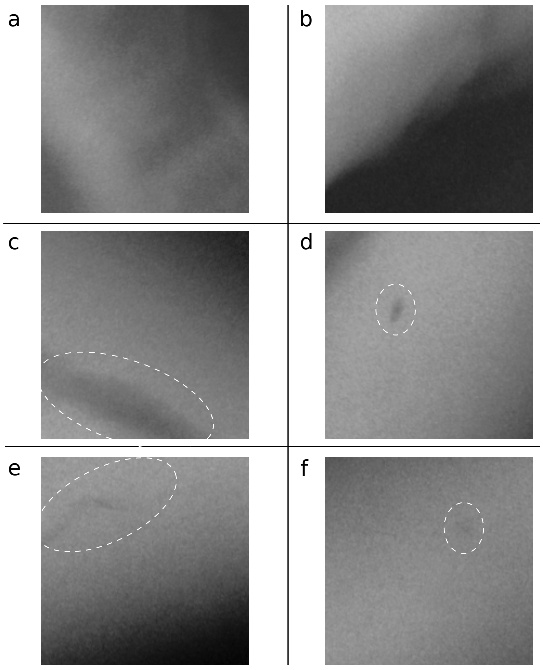

Figure 3: Example of weld patches, defects are encircled in dashed ellipses: a) and b) training set, no defects, accepted weld; c) and d) high contrast defects; e) mid contrast defects; f) low contrast defects. Reused from [2] with permission from ASME.

In Publication C the perturbation dataset was extended with synthetic indications. The synthetic indications were derived both from simple analytical models (see Fig. 8), random sampled with distributions on their parameters, as well as from visual images from outside of the application area. The latter, which we called natural image indications, was sampled as sub-regions from visual images (e.g. planes and birds), transformed to X-ray-like information with simple mathematical operations, before included into real X-ray images. An example of such a natural image indication can be seen in Fig. 7. The idea was that even though some of the natural image indications where unrealistic, that would not have a negative impact on the training, as long as some of the indications were realistic; and the approach to generating such indications were low on resources.

In Publication D, a similar approach to the visual inspection perturbation dataset was utilized. A representative perturbation dataset patch can be seen in Fig. 9.

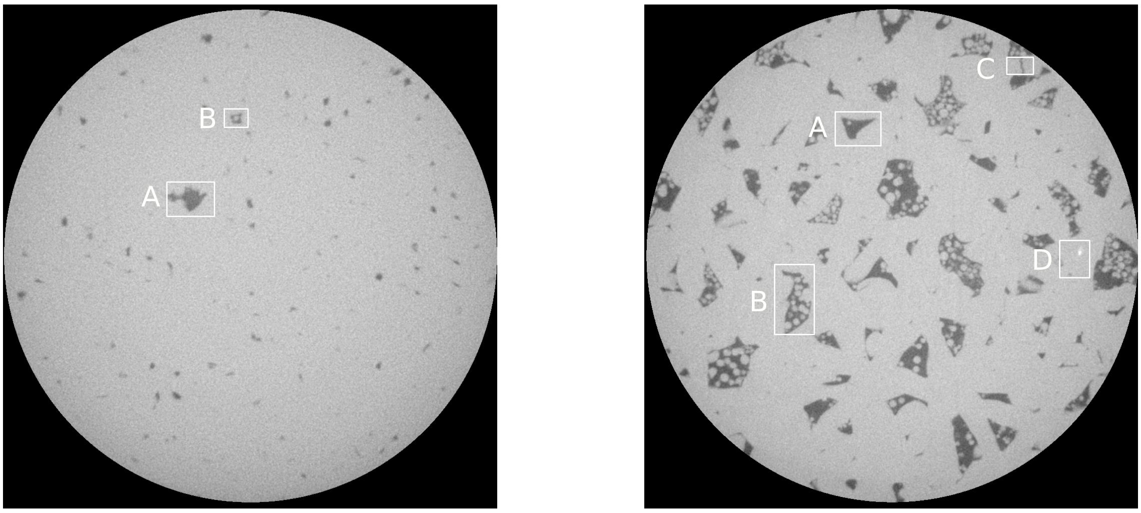

Figure 4: Example CT slices from the dataset [35], about $1000 \times 1000\,\mathrm{px}^2$ at a about $2.5\,\mu$m voxel (reconstruction volume element) size. Labelled in the slices are some different example indication types. Reused from [3].

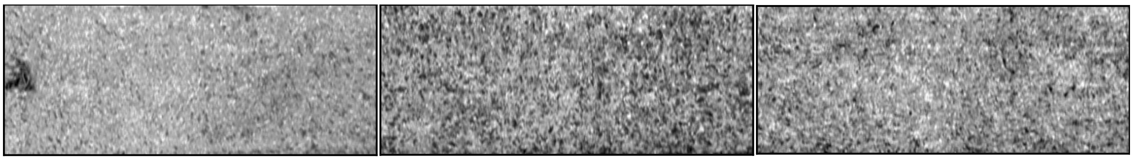

Figure 5: Examples of defect free samples. Reused from [5].

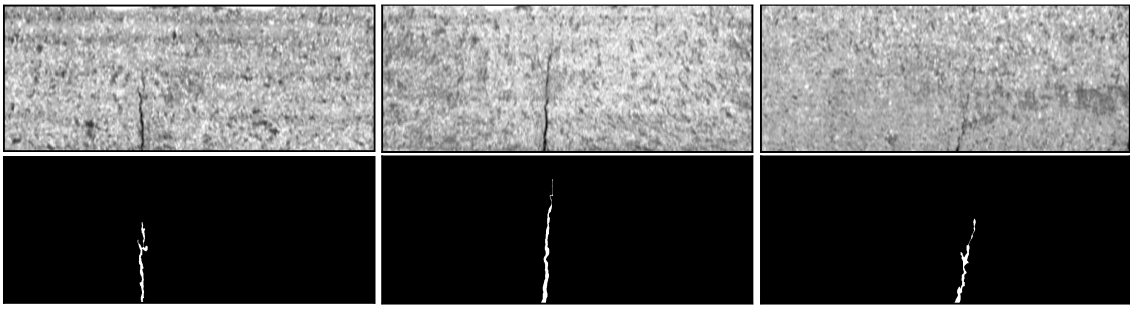

Figure 6: Examples of defect samples, the ground truth defect indication segmentations are indicated in the bottom row. Reused from [5].

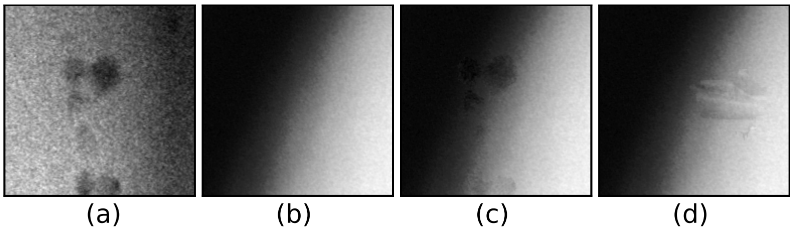

Figure 7: Examples of perturbed defect patches in the training dataset. (a) real defect, (b) ok weld, (c) ok plus real defect patch, and (d) ok weld plus a synthetic natural images based defect patch. Reused from [5].

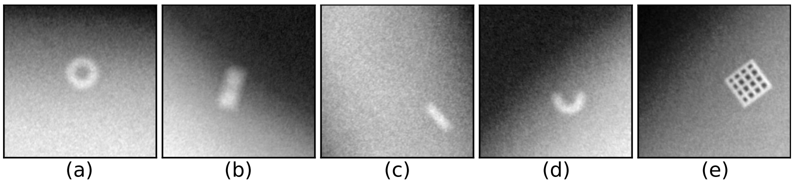

Figure 8: Examples of synthetic defect and anomaly indications inserted into weld ok patches: (a) circular hollow inclusion, (b) dogbone inclusion, (c) elongated inclusion, (d) partial circle inclusion, and (e) raster. Reused from [5].

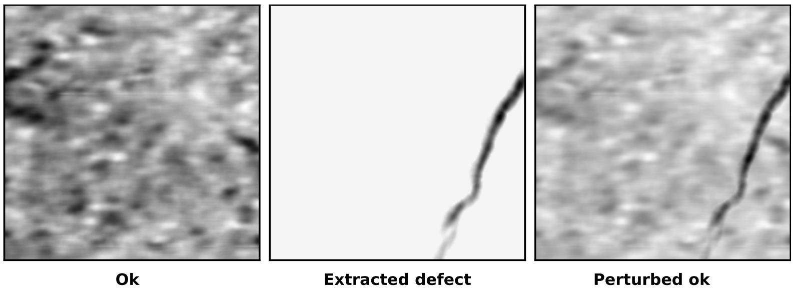

Figure 9: A representative perturbation dataset patch ($192 \times 192\,\mathrm{px}^2$) sample example. Reused from [5].

### 3.1.4 Results

In this section a small subset of the results will be given, see Publications A-D for more results and details. Overall, in all of the publications, our results indicated that the proposed AE models could successfully remove large amounts of accepted intensity variation in the images, and at the same time leaving potential anomalies to the residual image. For the X-ray-based images applications we showed that we could model the accepted variations down to the X-ray quantum noise levels.

The first results on the GDXray [34] dataset, from Publication A, showed that a high true positive rate at some $80-90\,\%$ could be achieved at false positive rates (about $4\,\%$ or less), for defects imaged at high to mid contrast to noise (see Fig. 3). These results were then improved in Publication C, with true positive rates at $90-94\,\%$ at false positive rates about $0.1\,\%$. The result improvement was most likely due to the improved perturbation dataset, extended with synthetic data; especially the synthetic natural indications dataset.

In Fig. 10 an example of the results for a mid-contrast real defect in the test dataset can be seen. Also, in Fig. 11 the results for synthetic material defect in the test dataset is shown.

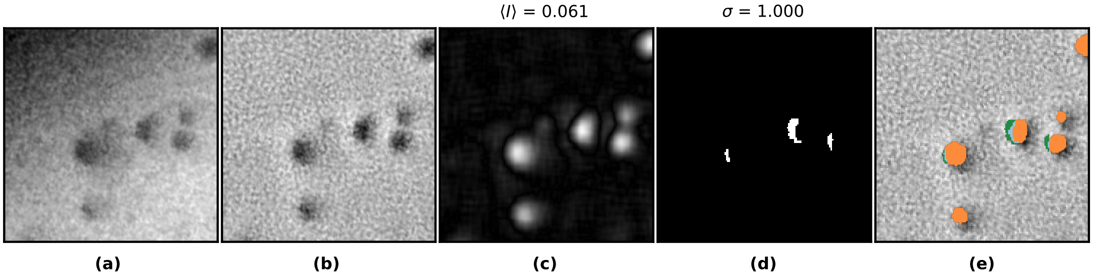

Figure 10: Results for the mid contrast real defects test dataset. (a) is input, (b) is residual image, residuals analysis kernels results are for average kernel in (c) and standard deviation in (d). Above the patch is the maximum value indicated. Kernels above thresholds are indicated in (e) with blue ($\langle I \rangle$ and $\sigma$), green ($\sigma$), and orange ($\langle I \rangle$). Reused from [4].

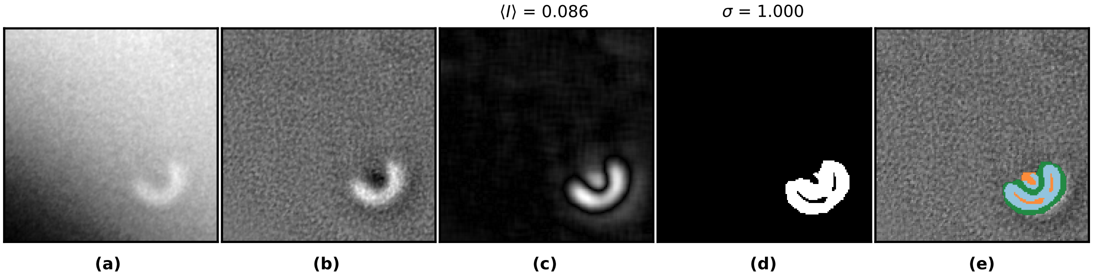

Figure 11: Results for the synthetic partial circle inclusion test dataset. (a) is input, (b) is residual image, residuals analysis kernels results are for average kernel in (c) and standard deviation in (d). Above the patch is the maximum value indicated. Kernels above thresholds are indicated in (e) with blue ($\langle I \rangle$ and $\sigma$), green ($\sigma$), and orange ($\langle I \rangle$). Reused from [4].

The model was also evaluated on a synthetic highly hypothetical OOD data example, far from the training dataset, the results can be seen in Fig. 12.

The result of the sliding window approach (sliding the smaller patch window over a larger input image, at each place applying the AE model) can be seen in Fig. 13. Also note that the model is not intended to be utilized for high precision segmentation, but rather it is intended to classify on a patch-level if the patch contain only accepted weld material or not; however as illustrated in the figure it offers some rough localization of where it did not meat this criteria.

Compared to the supervisedly trained patch level binary classifier, our approach showed less spread in OOD detection performance for different model individuals (independently trained model realizations), when trained with the natural synthetic indications perturbations. For the different synthetic OOD test cases, our proposed $\delta$AE model was either higher in true positive

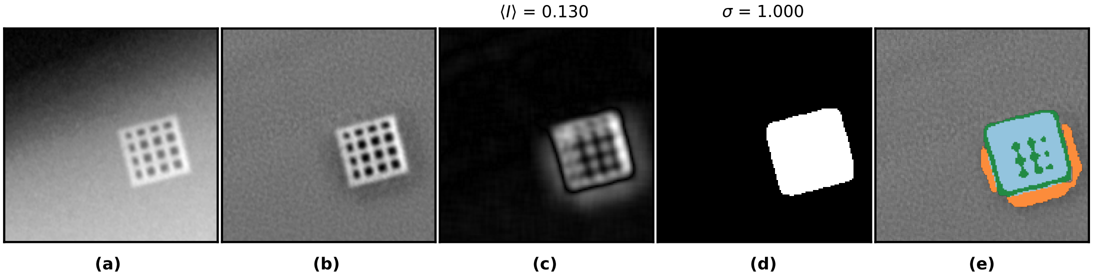

Figure 12: Results for the exotic synthetic raster anomaly. (a) is input, (b) is residual image, residuals analysis kernels results are for average kernel in (c) and standard deviation in (d). Above the patch is the maximum value indicated. Kernels above thresholds are indicated in (e) with blue ($\langle I \rangle$ and $\sigma$), green ($\sigma$), and orange ($\langle I \rangle$). Reused from [4].

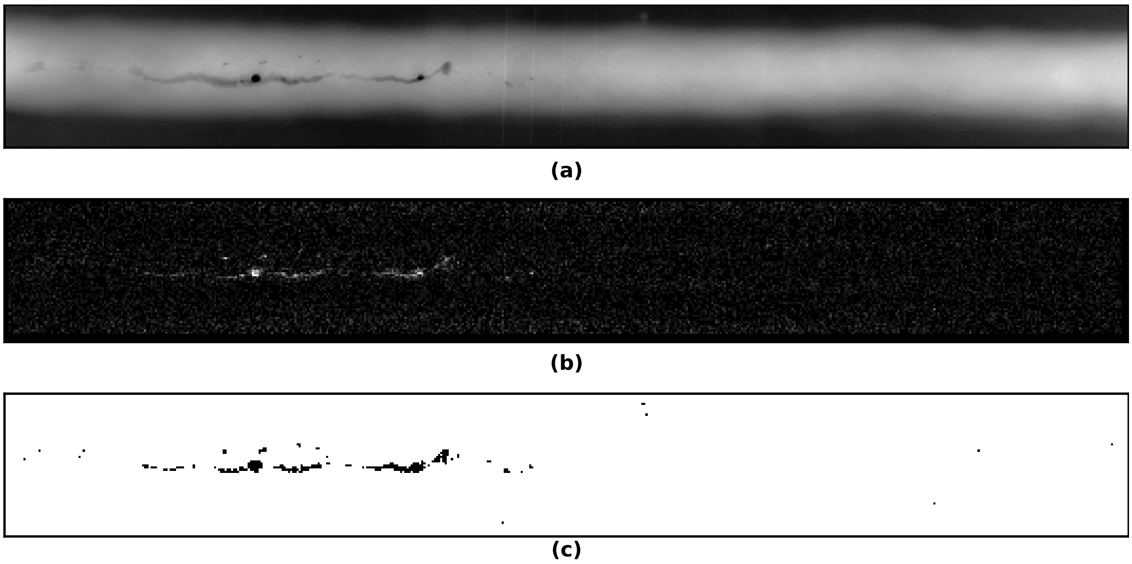

Figure 13: AE-model sliding window results for the test dataset. Trained with the synthetic natural indication perturbation dataset. (a) is the original input, (b) is the residuals, and (c) is the kernel analysis results thresholded with thresholds resulting in an FPR at 0.1 % on a patch level. Reused from [4].

rate or as good as the supervisedly trained model. When trained only with real experimental data in the perturbation dataset, the $\delta$AE was superior in OOD detection compared to the supervisedly trained models. For example, the OOD test data in 12, was only correctly classified in average about 70 % of the cases by the supervisedly trained classifier, when trained only on the real defect data; compared to the $\delta$AE which already without any perturbation dataset detected all of them correctly (at 0.1 % FPR).

Next, the results from Publication B, XCT application, will be summarized. A perturbation dataset consisting of modelled X-ray quantum noise and synthetic anomalies was derived and utilized. Overall a high TPR at 100 % was achieved on real inclusion-like test data (not similar to the training data) at an FPR at 1 %. For a synthetic hypothetical exotic machine element indication (see Fig. 14), with a contrast to noise ratio around 11, the TPR was 80 % at a FPR of 1 %.

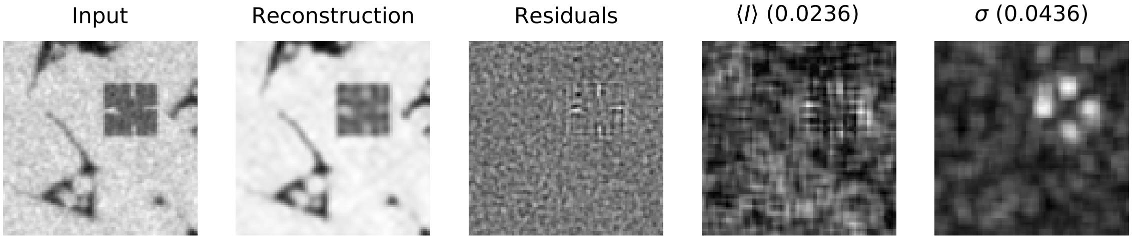

Figure 14: XCT results, X-ray noise during training, test image, with a synthetic exotic hypothetical machine element-like anomaly, $CNR \approx 11$. Thresholds set to FPR 1 %. Analysis kernel results are at the detection limit for the TPR at 1 %. Reused from [3].

The distribution of the kernel analysis values ($\langle I \rangle$, $\sigma$) is indicated in Fig. 15. The separation between training, test ok, and test anomaly datasets is visualized. Note that the train and test ok dataset subset contains more samples compared to the test anomaly (only 24); where the test anomaly are the real high density anomaly patches.

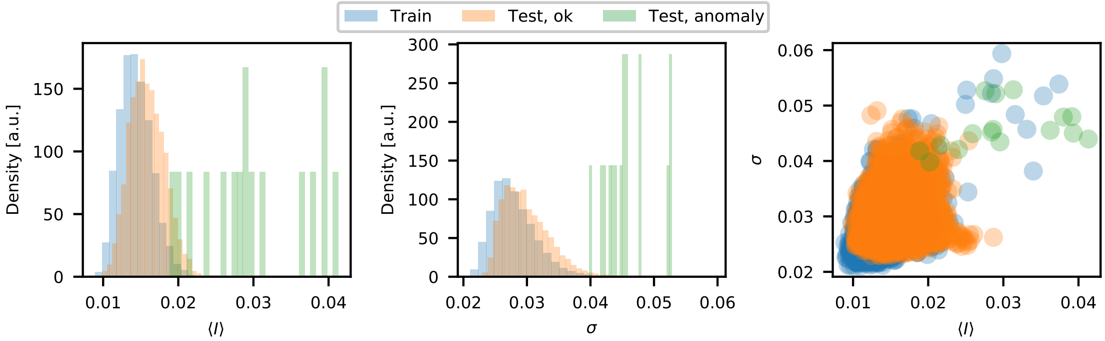

Figure 15: XCT OOD detection results, with X-ray noise active during training, with total loss $L+L_{maxmin}$. The residual kernel average $\langle I \rangle$ and standard deviation $\sigma$ are the max absolute values in each patch. The test anomaly is the real anomalies. Only a random subset of the training and test datasets are plotted. The density represents the probability density, with its sum over all histogram bins normalized to 1. Reused from [3].

The models were also evaluated on both pure noise patches (no material imperfection indication present) and a small validation dataset (held out from the training dataset) with patches containing material imperfection indications. The results in Fig. 16 together with the above results lead us to conclude that the proposed model can model the information of interest down to X-ray noise levels, and still fail to do so for OOD data (qualitatively) considered anomalies.

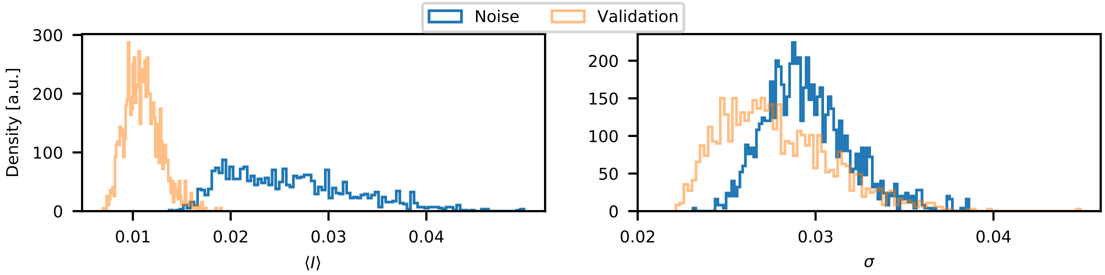

Figure 16: XCT Residual kernel (average $\langle I \rangle$ and standard deviation $\sigma$) distributions (given as probability density) evaluated over each patch (max value in each patch), for the pure noise case as well as the validation set; where the validation set is a held-out small part of the training dataset utilized during the training. Reused from [3].

As for the visual inspection application in Publication D, some representative results can be seen in Fig. 17. The best results were achieved with the low compression architectures with a TPR 90 %, FNR 99 %, FN 5 and FP 4. Which can be compared to previous studies with supervisedly trained models which achieved FP = 0 at FN = 0. Compared to earlier studies on unsupervisedly trained models, we achieved much better performance; however, as already pointed out, our $\delta$AE model can be argued to be supervisedly trained with intrinsic capability to detect OOD data at inference, rather than unsupervisedly trained.

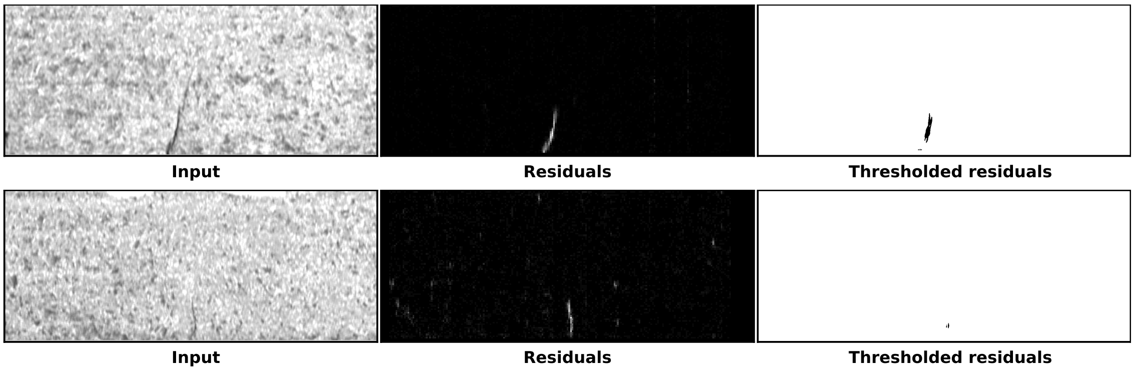

Figure 17: Example of sliding window results for input images from the test dataset with defects present and detected as such, examples of true positives. First row is a large (1044 pixel count) high contrast indication, and the second row is an example of the detection limit (at the settings chosen for TNR and FPR) with a small (53 pixel count) low contrast indication. Reused from [5].

The model was also evaluated successfully on a few synthetic OOD data examples, see Fig. 18.

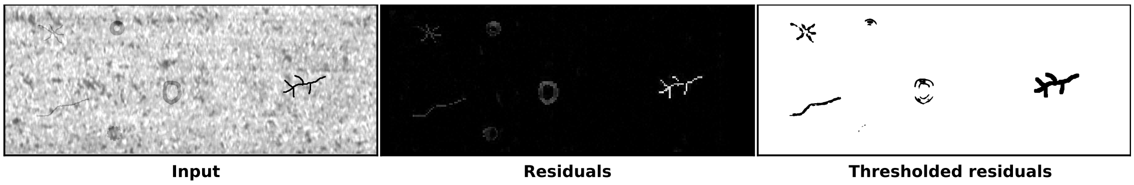

Figure 18: Example of sliding window results for the model in setup 5 on an input image from the test dataset with synthetic unexpected indications present, OOD examples. Six OOD indications are present, with one of them below the detection limits. Reused from [5].

## 3.2 Generative models (publication B)

In addition to exploring the problem of OOD data detection, or robust confidence estimation with respect to OOD data, we also explored Deep Learning-based generative models. The reason for this was that there is an intrinsic connection between the OOD detection problem (check if the input is similar to the training distribution) and the generative model problem (draw new unique samples similar to the training distribution).

Both generative models and OOD detection implicitly model the input distribution. While the OOD detection method will need to be forced or regularized not to model too much (despite inherently being a good generalizer), the opposite is in general true for the generative model; it will intrinsically generate images with inherently as low variation as possible, unless forced not to do so (a phenomenon known as "mode collapse").

### 3.2.1 Model architecture, training, and datasets

The explored generative model was a deep convolutional generative adversarial network (DC-GAN). A GAN [36] consists of two neural networks, one generative model generating new samples and one discriminative network trying to discern generated from real samples. They are trained at the same time, competing with each other.

An illustration of the generative part of the DCGAN is show in Fig. 19. The DCGAN architecture is made similar to the AE model used in this project. The input ($G_0$) to the generator consist of normal distributed random variables. It is fed through three fully connected layers ($G_D$). Then follows a series of three layers ($G_{C1}$ to $G_{C3}$) consisting of 2D convolutions, similar to the decoder part in the AE model. The discriminator part of the DCGAN is make symmetric to the generator. See [3] for details.

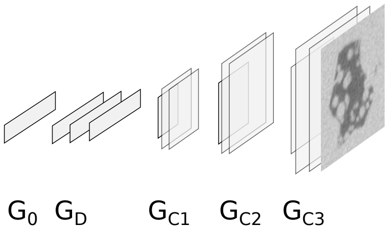

Figure 19: Illustration of the DCGAN. The discriminator network for the GAN is similar to the generator network but with the direction reversed. Reused from [3].

Despite the practical issues with training GANs, they are known to generate realistic new samples. However, generating just a few realistic samples is not enough for our envisioned industrial applications; rather we want to generate a realistic sample distribution, covering as much aspects of the training distribution as possible. Therefore, mode collapse mitigation approaches were explored for the generator loss, more specifically feature matching [37].

In earlier studies on training GANs for generating X-ray images the generated images have been more or less full of artefacts. To mitigate the formation of artefacts we explored the effects of adding X-ray-like noise, with simple known analytical physics-based mathematical models, to the images. The noise was added to the DCGAN generated images, prior to being fed into the discriminating model. In a sense, ideally, training the DCGAN to generate images free of X-ray-noise and then add the noise at a suitable level just after the DCGAN. The same XCT dataset [35] as utilized for the OOD detector was utilized also for the DCGAN experiments.

### 3.2.2 Results

As can be seen in Fig. 20 and 21, adding X-ray-noise with the analytical model resulted in more realistic looking generated X-ray images and the amount of artefacts was considerably decreased.

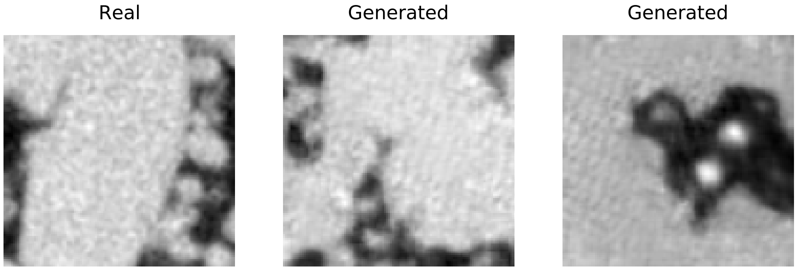

Figure 20: A patch sample (left) compared to two (middle, right) DCGAN generated (fake) material imperfection, with no noise added during the training. Reused from [3].

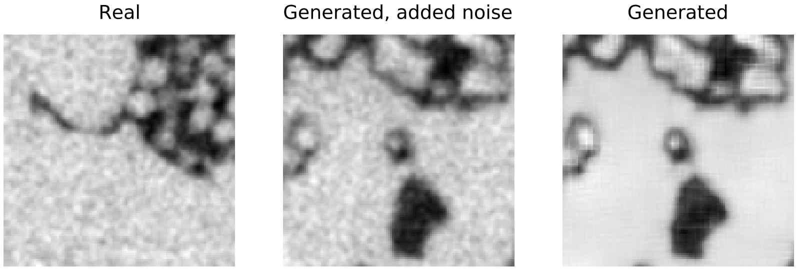

Figure 21: A real patch sample compared to a DCGAN generated (fake) material imperfection, with X-ray noise added during the training. Reused from [3].

Highly realistic imperfection indications could be generated, with considerable variation between generated samples. Some representative examples are shown in Fig. 22 and 23. Also, a model trained on a class of imperfections with low amount of internal voids (Class A) did generate samples considerable dissimilar to another class defined as imperfections with high amount of internal voids (Class B); compare differences between Fig. 22 and 23.

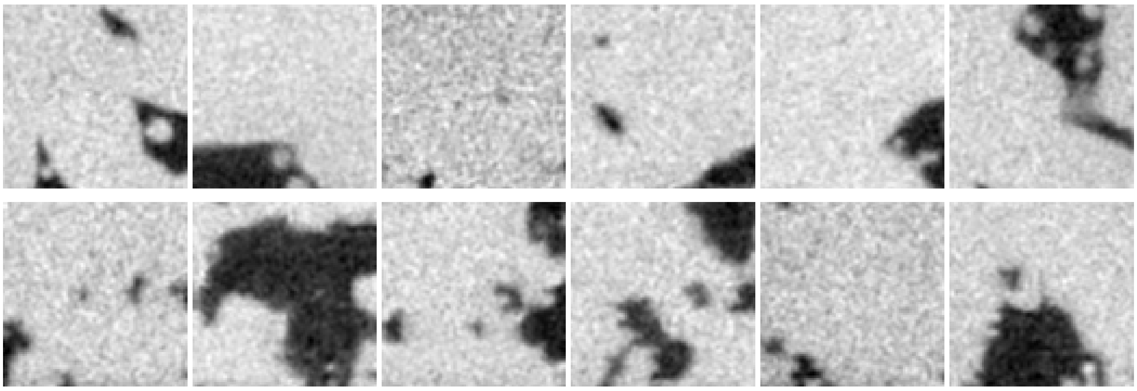

Figure 22: Top row are real and bottom row are GAN generated (fake) material imperfection indications with X-ray noise added during the training. Trained on the class of imperfections with low amount of inner voids. Reused from [3].

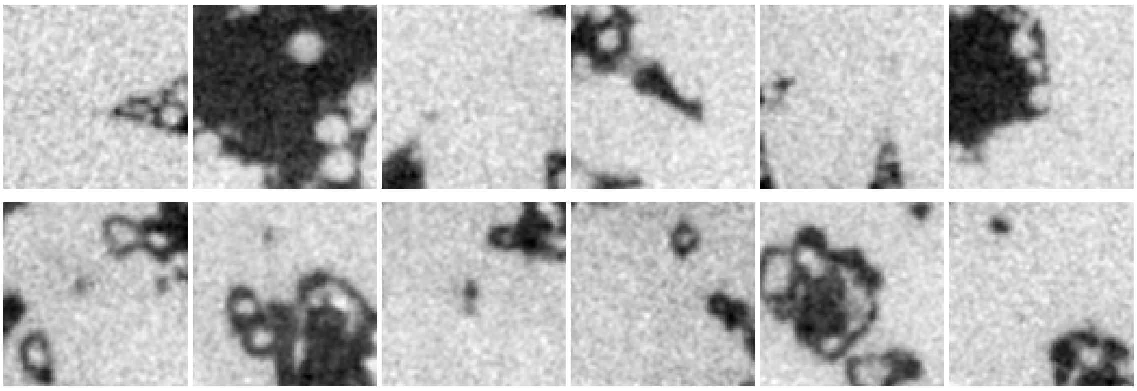

Figure 23: Top row are real and bottom row are GAN generated (fake) material imperfection indications with X-ray noise added during the training. Trained on the class of imperfections with high amount of inner voids. Reused from [3].

However, we question if the generative models really model the whole training distribution. For example, though we did not make that experiment, we suspect that a human still can discern, with a high probability, the generated (fake) images from real ones if shown a small set of images at a time rather than a single image at a time.

# 4 Conclusions and future research

In this project we have derived and explored a Deep Learning-based out-of-distribution (OOD) detector model which is capable of both correctly classifying input data similar to the training data, as well as to detect OOD data examples. The focus application was X-ray-based inspection methods, i.e. industrial X-ray images in 2D and 3D. However we also successfully generalized the model to visual inspection data.

The explored approach was based on training an Auto-Encoder to model (reconstruct) a given accepted image distribution, but at the same time reject to model added perturbations. The perturbations representing OOD data, highly structural noise, for example material imperfections in welds or additively manufactured metals, or anything outside of the accepted distribution. Input which the model then fails to model (reconstruct) can be considered OOD data. We call the model a Perturbed Auto-Encoder ($\delta$AE).

We argue that the proposed approach, which can be considered somewhere between supervised and unsupervised methods, has intrinsically built-in a more sensible reaction to OOD data at inference time, compared to conventionally classifiers trained supervisedly (e.g. binary segmentation models). This aspect we also demonstrated in the project. The performance (on in-distribution test data) of our proposed perturbed AE was in many comparisons much better than unsupervisedly trained OOD approaches, and sometimes a bit lower than or similar in performance to the conventional supervisdely trained models. For detecting OOD data it however was much better than the conventionally supervisedly trained models.

In connection to the OOD detector we also explored similar architectures for generative models. A so called Generative Adversarial Network (GAN) was trained and artefacts were shown to decrease by adding X-ray quantum noise explicitly; in effect including what we already know about the imaging system, rather than just trying to model the image data as any image. Similar considerations of the specific application's pros and cons, the X-ray-imaging physics, were shown to increase also the performance of the OOD detector.

Essentially all DL models require large amounts of training data. Also, to label the data correctly with human labor is also resource demanding. In this project we successfully explored hybrid data, real data annotated with synthetic data. This both growing the datasets as well as decreasing the resources required for labeling. More specifically we demonstrated that it was possible to utilized synthetic data from completely different domains (to get a high variation) and just apply simple mathematical transformations to bring it into the X-ray-image domain. We showed that this could be utilized successful even though not all of the generated data looked realistic.

We propose that future research for these perturbed Auto-Encoder models could be towards more systematic approaches to deriving suitable perturbation datasets. Generative DL-based models could potentially be utilized as such perturbation datasets, as long as the variation within generated sample distribution is representative (large) enough. In this work we explored a GAN model, though other generative DL models could potentially be fruitful to explore also for the perturbation datasets. We also propose that the perturbed Auto-Encoder solution derived in this project should be evaluated in other applications domains, for example medical X-ray imaging could be reasonable.

# References

[1] Nobert G. Meyendorf, Leonard J. Bond, J. Curtis-Beard, S. Heilmann, Saveri Pal, R. Schallert, H. Scholz, and C. Wunderlich. Nde 4.0—nde for the 21st century—the internet of things and cyber physical systems will revolutionize nde. In *Proceedings of the 15th Asia Pacific Conference for Non-Destructive Testing (APCNDT 2017)*, 2017.

[2] Erik Lindgren and Christopher Zach. Autoencoder-based anomaly detection in industrial x-ray images. volume 2021 48th Annual Review of Progress in Quantitative Nondestructive Evaluation of *Quantitative Nondestructive Evaluation*, 2021.

[3] Erik Lindgren and Christopher Zach. Analysis of industrial x-ray computed tomography data with deep neural networks. In *Developments in X-Ray Tomography XIII*, volume 11840, page 118400B. International Society for Optics and Photonics, SPIE, 2021.

[4] Erik Lindgren and Christopher Zach. Industrial x-ray image analysis with deep neural networks robust to unexpected input data. *Metals*, 12(11), 2022.

[5] Erik Lindgren and Christopher Zach. Deep-learning-based out-of-distribution data detection in visual inspection images. In Norbert G. Meyendorf, Christopher Niezrecki, and Ripi Singh, editors, *NDE 4.0, Predictive Maintenance, Communication, and Energy Systems: The Digital Transformation of NDE*, volume 12489, page 1248909. International Society for Optics and Photonics, SPIE, 2023.

[6] V.R. Rathod and R.S. Anand. A comparative study of different segmentation techniques for detection of flaws in nde weld images. *J Nondestruct Eval*, 31:1–16, 2011.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* The MIT Press, 2016.

[8] Valérie Kaftandjian, Olivier Dupuis, Daniel Babot, and Yue Min Zhu. Uncertainty modelling using dempster–shafer theory for improving detection of weld defects. *Pattern Recognition Letters*, 24(1):547–564, 2003.

[9] V Lashkia. Defect detection in x-ray images using fuzzy reasoning. *Image and Vision Computing*, 19(5):261–269, 2001.

[10] Yan Wang, Yi Sun, Peng Lv, and Hao Wang. Detection of line weld defects based on multiple thresholds and support vector machine. *NDT & E International*, 41(7):517–524, 2008.

[11] Rafael Vilar, Juan Zapata, and Ramon Ruiz. An automatic system of classification of weld defects in radiographic images. *NDT & E International*, 42(5):467–476, 2009.

[12] J. Kumar, R.S. Anand, and S.P. Srivastava. Flaws classification using ann for radiographic weld images. In *International Conference on Signal Processing and Integrated Networks*, pages 145–150, 2014.

[13] X. Dong, C. J. Taylor, and T. F. Cootes. Automatic inspection of aerospace welds using x-ray images. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2002–2007, 2018.

[14] Wenhui Hou, Ye Wei, Jie Guo, Yi Jin, and Chang'an Zhu. Automatic detection of welding defects using deep neural network. *Journal of Physics: Conference Series*, 933:012006, jan 2018.

[15] Wenhui Hou, Ye Wei, Yi Jin, and Changan Zhu. Deep features based on a dcnn model for classifying imbalanced weld flaw types. *Measurement*, 131:482 – 489, 2019.

[16] L. Yang and H. Jiang. Weld defect classification in radiographic images using unified deep neural network with multi-level features. *J Intell Manuf*, 32:459–469, 2021.

[17] Roger Booto Tokime, Xavier Maldague, and Luc Perron. Automatic defect detection for x-ray inspection: Semantic segmentation with deep convolutional network. In *International Industrial Radiology and Computed Tomography DIR2019*, 2019.

[18] Topias Tyystjärvi, Iikka Virkkunen, Peter Fridolf, Anders Rosell, and Zuheir Barsoum. Automated defect detection in digital radiography of aerospace welds using deep learning. *Welding in the World*, 66:643–671, 2022.

[19] Domingo Mery. Aluminum casting inspection using deep learning: A method based on convolutional neural networks. *Journal of Nondestructive Evaluation*, 39:12, 2020.

[20] Patrick Fuchs, Thorben Kröger, Tobias Dierig, and Christoph Garbe. Generating meaningful synthetic ground truth for pore detection in cast aluminium parts. In *9th Conference on Industrial Computed Tomography, Padova, Italy, iCT2019*, 2019.

[21] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31:759–779, 2020.

[22] Gaokai Liu, Ning Yang, Lei Guo, Shiping Guo, and Zhi Chen. A one-stage approach for surface anomaly detection with background suppression strategies. *Sensors*, 20(7), 2020.

[23] Shuang Mei, Jiangtao Cheng, Xin He, Hao Hu, and Guojun Wen. A novel weakly supervised ensemble learning framework for automated pixel-wise industry anomaly detection. *IEEE Sensors Journal*, 22(2):1560–1570, 2022.

[24] Zhongqin Bi, Qiancong Wu, Meijing Shan, and Wei Zhong. Segmentation-based decision networks for steel surface defect detection. *Journal of Internet Technology*, 23(6):1405–1416, 2022.

[25] Rémi Cogranne and Florent Retraint. Statistical detection of defects in radiographic images using an adaptive parametric model. *Signal Processing*, 96:173 – 189, 2014.

[26] Robert Grandin and Joe Gray. Implementation of automated 3d defect detection for low signa-to noise features in nde data. In *AIP Conference Proceedings 1581*, 2014.

[27] I.G. Kazantsev, I. Lemahieu, G.I. Salov, and R. Denys. Statistical detection of defects in radiographic images in nondestructive testing. *Signal Processing*, 82(5):791 – 801, 2002.

[28] I. Tošić and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.

[29] Alice Presenti, Zhihua Liang, Luis Filipe Alves Pereira, Jan Sijbers, and Jan De Beenhouwer. Automatic anomaly detection from x-ray images based on autoencoder. *Nondestructive Testing and Evaluation*, 2022.

[30] Weitao Tang, Corey M. Vian, Ziyang Tang, and Baijian Yang. Anomaly detection of core failures in die casting x-ray inspection images using a convolutional autoencoder. *Machine Vision and Applications*, 32:102, 2021.

[31] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

[32] Justus Zipfel, Felix Verworner, Marco Fischer, Uwe Wieland, Mathias Kraus, and Patrick Zschech. Anomaly detection for industrial quality assurance: A comparative evaluation of unsupervised deep learning models. *Computers & Industrial Engineering*, 177:109045, 2023.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[34] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34:42, 2015.

[35] F.H. Kim, S.P. Moylan, E.J. Garboczi, and J.A. Slotwinski. Investigation of pore structure in cobalt chrome additivelymanufactured parts using x-ray computed tomography andthree-dimensional image analysis. *Additive Manufacturing*, 17:23–38, 2017.

[36] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2234–2242, Barcelona, Spain, 2016.