WPEDU-CPR(18) 19179

R Skilton et al.

Unsupervised Learning of Generative Models for Robotic Visual Anomaly Detection

Preprint of Paper to be submitted for publication in Proceeding of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Unsupervised Learning of Generative Models for Robotic Visual Anomaly Detection

Robert Skilton[1] and Yang Gao[2]

***Abstract*— Visual anomaly detection is an important task for robotic maintenance systems, which are employed in hazardous environments such as nuclear reactors. Generative Adversarial Nets (GANs) are a popular method for creating models that are able to generate complex data samples such as natural images. Recent research has demonstrated their capability in image in-painting and medical anomaly detection, by searching for the latent space representation, from which similar samples can be generated. This is, however significantly computationally intensive and operates on timescales that are prohibitive for real-time applications. This paper proposes a method for automated anomaly detection in images, using a regeneration method which is compatible with real-time applications. The method for image regeneration is shown to work at timescales appropriate for real-time applications whilst suffering only a small loss in accuracy over previous techniques. An anomaly detection method which calculates residuals between an image and its corresponding regeneration is proposed and evaluated.**

I. Introduction

Visual anomaly detection relates to the problem of using visual information, i.e. images or sequences of images from a camera system, to detect anomalies which may be damaged or defective items, or items which were not anticipated to be present. In our particular application case - a Tokamak experimental nuclear fusion reactor, we are interested in both; detecting damaged components (for an example see Figure 1), as well as detecting objects that may have been accidentally left behind following a maintenance campaign.

Generative models have been demonstrated to be capable of regenerating whole or partial data structures such as images, using semantic information to accurately interpolate and extrapolate highly complex and nonlinear information. This type of image reconstruction is applicable to many real-world problems including anomaly detection in industrial and medical applications, and robot situational awareness. Prediction of future states is widely seen to be a key component of computational intelligence. Generative Models such as Variational AutoEncoders [1] and Generative Adversarial Nets [2] may form a starting point for such a predictive intelligence through their ability to predict likely input data.

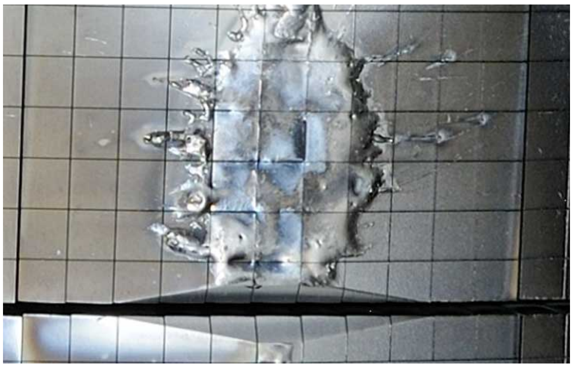

Fig. 1: Within nuclear fusion facilities, as well as a number of other industrial application domains, it is important to detect anomalous items using visual information. The image shows a tile which has been damaged by being splashed with liquid metal.

Generative Adversarial Nets learn to generate data samples that are semantically similar to training samples with high syntactic correctness, whilst not actually directly learning training data. As such they are able to generate entirely novel samples.

GANs have also been demonstrated to have the ability to reconstruct partial query samples by searching for the latent distribution that will generate a sample that is similar to the query sample. This effectively demonstrates that trained generative models of this type are able to capture complex patterns of what is "expected" within the input data. As such, this type of generative model is useful not only for filling in missing information, but can also be used as a measure of expectedness of parts of the input data. By

*This work was supported by EUROfusion

[1] Robert Skilton is with Remote Applications in Challenging Environments (RACE), UK Atomic Energy Authority, Culham Science Centre, Oxfordshire, UK. Robert is also with the STAR lab at Surrey. robert.skilton@ukaea.uk

[2] Yang Gao is head of the Surrey Technology for Autonomous systems and Robotics (STAR) Lab, University of Surrey, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, UK yang.gao@surrey.ac.uk

learning to generate what is normal, novelty can be detected by identifying elements that have not been generated, through measuring the residual between the generated data sample, and the original query sample.

In this paper we demonstrate a technique which is able to detect anomalous or novel items in images, using an approach that uses generative black-box modelling, trained using an unsupervised training method, to learn what "normal" data samples or images should look like. A method for generating normal samples close to a given query image is presented, allowing generated "normal" data samples to be compared with query data for novel information. The method is therefore thought to be capable of detecting novel items without any a priori information on the nature of those new or novel items.

II. Related Work

Generative Adversarial Networks [2] are a recent approach to training generative models based on an unsupervised, adversarial approach. They have been demonstrated to have capability in generating complex natural images [3].

Generative adversarial networks consist of two main elements, the Generator G, and a Discriminator D. The role of the discriminator is to estimate the probability that a given data sample (e.g. an image) is a natural image as opposed to an artificial generated image. The role of the generator is to attempt to randomly generate realistic data samples that are able to fool the discriminator. As such, the Generator and the Discriminator are playing adversarial roles in a 2-player game, which can be described in the minimax function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}}[log(D(x))] + \mathbb{E}_{z\sim P_{z(z)}}[log(1-D(G(z)))] \tag{1}$$

The Generator G operates on a simple random prior, $z \sim p_z$, and implicitly defines a probability distribution p_g of generated samples $G(z)$. In order for the generated samples to match the real-world data, it is desrable for p_g to converge to p_{data}, the distribution of natural training samples.

With $D(x)$, the output of the Discriminator being defined as the probability that sample x came from the data (as opposed to being generated), the optimisation problem can be defined as back-propagation by ascending the gradient:

$$\nabla_{\theta d} \frac{1}{m} \sum_{i=1}^{m} [\log(D(x^{(i)})) + \log(1-D(G(z^{(i)})))] \tag{2}$$

and descending the gradient:

$$\nabla_{\theta g} \frac{1}{m} \sum_{i=1}^{m} \log(1-D(G(z^{(i)}))) \tag{3}$$

In other words, maximising the log probability that $x^{(i)}$ is marked as coming from the training data plus the log probability that $z^{(i)}$ is flagged as being generated by adjusting the Discriminator parameters, whilst simultaneously minimising the log probability that $z^{(i)}$ is marked as being artificially generated by adjusting the Generator parameters.

Radford et al. [4] and Salimans et al. [5] build substantially on this foundation by proposing a number of improvements to the architectures and training processes, as well as providing means for assessing the performance of GANs, and providing insights into the types of representations learned. It is made clear that continuous, linearly interpolatable, semantic representations are implicitly mapped from the latent space, and this is demonstrated qualitatively through interpolation and vector arithmetic in latent space, generated into image space representations which semantically follow the interpolation or arithmetic.

Semantic Image In-painting with Deep Generative Models [6] attempts to make use of GANs to complete partial images with generated photo-realistic data. GANs have been shown to be capable of generating high-quality, realistic images [4], however completion of partial images required conditioning and constraining on the context provided by the contents of the partial image. The Generator once trained on a data set would normally generate random images which would likely be significantly different from the partial query image.

In order to achieve this, an attempt is made to find the encoding in latent space $\hat{z}$ which most closely relates to the partial image data x in image space. Once this is achieved, the image data can be reconstructed using the generated image resulting from this encoding, $G(\hat{z})$.

With this method and architecture, the authors were able to generate strikingly impressive results which, qualitatively look highly realistic, yet performed relatively poorly on standardised signal-to-noise ratio based metrics such as Peak Signal-to-Noise Ratio (PSNR). This is attributed to the networks generating semantically, but not, necessarily visually similar images, questioning validity of the standard metrics.

Generative Adversarial Networks have been used in [7] for anomaly detection in medical imaging data. A GAN was first trained on healthy samples, and then used to predict anomalies based on methods in [6] for finding the closest Generated data to the real, query data.

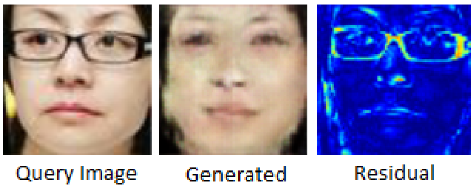

Fig. 2: GANs can be used to regenerate images based on learned models of expected contents. Generated images can therefore be compared with actual input images in order to automatically identify regions of novel information that has not been learned by the model during training.

Anomalies are then detected by adapting the coefficients of the latent distribution from which images are generated (z) by backpropagation, and an anomaly score $A(x)$ is produced, which can be used for detection of anomalous regions within an image. The final residual image can be used to identify anomalous regions.

Another significant disadvantage is that the input regeneration process requires optimising via backpropagation which is a significantly time-consuming process. Bi-directional GANs [8], and the ALI model [9] provide another method for learning to map input samples to latent space using a third element, an Encoder, which is solely responsible for this task. We utilise a simpler method, a latent regressor which is also able to perform this task.

This paper applies GAN models, with the capability of predicting latent representations to the task of image regeneration and assess the performance. This is undertaken with the eventual goal of robotic anomaly detection in mind.

III. Image Regeneration for Anomaly Detection

We hypothesise that by using a generative model trained on normal data to create samples that are close to the input image and then comparing the 2, we can find novel information which the generative model is unable to recreate (See figure 2).

The generative model was trained using only normal data (with no known anoamlies), in order that the model could represent what is "normal". I.e. the model should not have any in-built representation of what could be classed as anomalous.

During training, the discriminator was simultaneously trained to not only discriminate between real data and synthetic data samples produced by the generator, but also to predict the latent representation that the image came from (assuming the image was created by the Generator).

After training, the generator and discriminator could then be inverted (see Figure 4) in order to sequentially predict a latent representation $\hat{z}$ corresponding to the query image, and then, feeding this in as the input to the Generator, attemt to regenerate the query image using the trained "normal" image generative model.

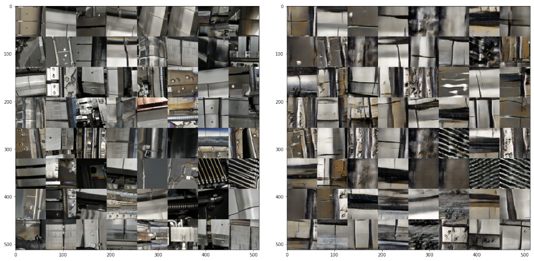

Fig. 3: Examples of normal (non-anomalous) examples from the image test set on the left, with the regenerated images on the right. The regenerated images are visually similar to the originals.

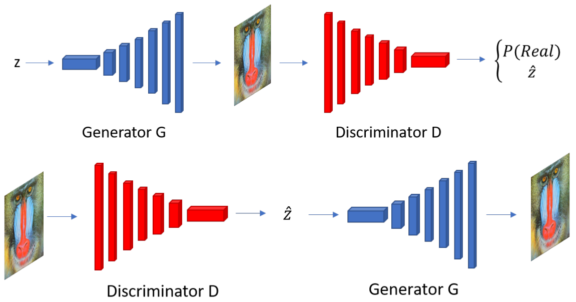

Fig. 4: Architecture of the GAN-based predictive regeneration network. At training time (top) the architecture is as with traditional GAN usage, with the additional output from the discriminator being a prediction of the latent representation, $\hat{z}$. At runtime (bottom), the architecture is reversed, with predicted latent space representations being used to seed the Generator.

Regenerations were carried out in this way for both normal images and images containing anomalies, whilst various methods for comparing the data samples were assessed. The methods used to assess the anomalousness of the input, based on the imput image in combination with other items relating to the generative model are as follows:

1) Residual scores (difference between input and regenerated)
2) Discriminator scores for input image
3) Discriminator scores for regenerated image
4) PSNR (Peak Signal-Noise Ratio) of regenerated image
5) SSIM (Structural Similarity) of regenerated image
6) Weighted sum of 1 and 3 (Similar to AnoGAN score)

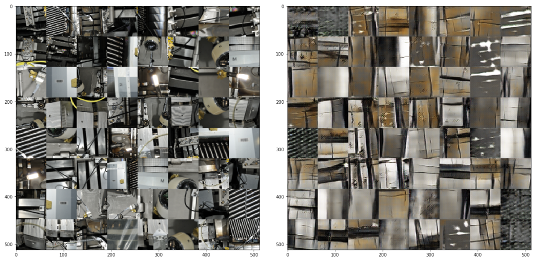

Fig. 5: Examples of Anomalous images (those containing maintenance equipment) on the left, and attempted regenerations on the right. The regenerations are noticeably different from the originals, in particular not featuring the novel remote maintenance equipment. Our hypothesis is that these discrepancies can be used to predict anomalies.

7) Discriminator score difference

This includes several comparitive methods, assessing difference between the input and regenerated image, as well as metrics that look at discriminator scores for input images and regenerated images individually. Figure 3 shows some examples of image regenerations and the resulting residual images.

A. Dataset

The experiments were conducted using a custom dataset relating to visible components within the Joint European Torus (JET) Tokamak, currently the worlds largest nuclear fusion energy experiment, located in Oxfordshire, UK.

The JET dataset consists of approximately 13,000 training images taken from photogrammetric surveys of the machine, all of which are 64x64 sized samples from larger images taken at various scales. Images contain views of various metallic components that form the experimental fusion reactor.

One subset of the images contains images of only the JET Tokamak machine. The other contains additional equipment used for maintenance. The practical applicability of this is in detecting items accidentally left behind during a maintenance campaign, however the technique is independent of the exact objects deemed to be anomalous, and any useful results should extend to any unforseen articles within the environment, such as other objects, missing components, and damaged components.

In addition, the Kinship Face in the Wild (KinFaceW-I) [10] dataset, was used as a secondary test case of regeneration performance.

IV. Architecture and Training

With standard Generative Adversarial Networks (GANs), the Generator and the Discriminator are playing adversarial roles in a 2-player game, which can be described in the minimax function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}}[log(D(x))] + \mathbb{E}_{z\sim P_{z(z)}}[log(1-D(G(z)))] \tag{4}$$

The Generator G operates on a simple random prior, $z \sim p_z$, and implicitly defines a probability distribution p_g of generated samples $G(z)$. In order for the generated samples to match the real-world data, it is desrable for p_g to converge to p_{data}, the distribution of natural training samples.

With $D(x)$, the output of the Discriminator being defined as the probability that sample x came from the data (as opposed to being generated), the optimisation problem can be defined as backpropagation by ascending the gradient:

$$\nabla_{\theta d}\frac{1}{m}\sum_{i=1}^{m}[\log(D(x^{(i)}))+\log(1-D(G(z^{(i)})))] \tag{5}$$

and descending the gradient:

$$\nabla_{\theta g}\frac{1}{m}\sum_{i=1}^{m}\log(1-D(G(z^{(i)}))) \tag{6}$$

I.e. maximising the log probability that $x^{(i)}$ is marked as coming from the training data plus the log probability that $z^{(i)}$ is flagged as being generated by adjusting the Discriminator parameters, whilst simultaneously minimising the log probability that $z^{(i)}$ is marked as being artificially generated by adjusting the Generator parameters.

Our first modification is to train the GAN not only to output the probability that the input, x, is a real image, but also output a prediction of the latent space representation that the image x came from, $\hat{z}$. Rather than arriving at this through gradient descent, as in [6], we would like the discriminator to learn this mapping during training.

The additional network output is parametrised as a vector of (100) real-valued numbers, output by a final linear layer in parallel with the logistic sigmoid layer for predicting image realism probability. The new output is generated using a single linear layer, which is trained using an L1 loss:

$$L_{z\_pred\_i} = |z_i - \hat{z}_i| \tag{7}$$

so the discriminator update step of the training program becomes ASCEND:

$$\nabla_{\theta d}\frac{1}{m}\sum_{i=1}^{m}[\log(D(x^{(i)})) + \log(1 - D(G(z^{(i)}))) \\ + |z_i - \hat{z}_i|] \tag{8}$$

Other generation methods and loss functions were tried including use of tanh and Logistic sigmoids to constrain $\hat{z}$ to the region of interest, and L2 and Cross Entropy loss functions, however the combination of linear outputs with L1 loss was found to work best in practice.

During training, the mapping from $z \rightarrow G(z)$ is constantly changing, and so L_{z_pred} is not guaranteed to converge, however L_{z_pred} does not affect the performance of G. One strategy for getting L_{z_pred} to converge would be to add a final training step, updating only the Discriminator, once good performance has been achieved in the Generator.

Training was carried out over 250 epochs of the 1000 image dataset, with randomly assigned initial network weights (no transfer learning). It is worth noting that although not quantified here, the addition of the z predictor to the discriminator network had little observed negative impact on training of the Generator.

Our implementation made use of the TensorFlow [11] library, and was trained using Tesla P100 GPUs on a set of non-anomalous images (See Figure 3).

V. Results

This section summarises results of the evaluations, firstly by examining the performance of the GAN-based image regeneration method, using standard metrics of image similarity, and secondly, looking at the performance of above mentioned methods for identifying anomalous images and anomalous contents within images.

A. Regeneration Performance

Regenerations were timed using a single CPU machine, with identical models and weights being used for each method. Time taken for the predictive model to regenerate a batch of 64 images was approximately 1 second, whilst elapsed time during the iterative process (1000 iterations) was 90 minutes (5400 seconds). We therefore report a non-trivial improvement in speed of image regeneration of the order of 5000 times.

Regenerated image accuracy is assessed using 2 standard metrics, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) [12]. Results are then compared between images generated using each of the methods.

TABLE I: Summary of PSNR metric results (in dB)

	PSNR (Backproapagated Z)	PSNR (Predicted Z)	Ratio
KinFaceW-I	21.36	15.64	0.78
JET	18.48	12.49	0.68

TABLE II: Summary of SSIM metric results

	SSIM (Backproapagated Z)	SSIM (Predicted Z)	Ratio
KinFaceW-I	0.74	0.55	0.74
JET	0.42	0.22	0.52

PSNR results show a mean PSNR of 15.64dB, compared to 21.36dB for the backpropagation method. This gives a PSNR ratio of 0.78. Mean SSIM index values are 0.55 and 0.74 for our method, and the backpropagation method respectively. SSIM index ratio is therefore 0.74.

A single score ratio can then be obtained by averaging the two score ratios, giving 0.76, a value which indicates overall performance.

Quantatative regeneration accuracy results are summarised and presented in Tables I and II.

B. Anomaly Detection Performance

Anomaly detection performance was assessed for each of 7 metrics, each using different aspects of the GAN system to try to predict anomalies.

Each of the identified methods assigns a single scalar anomalousness score to each image. For each metric score set, detection performance was tested across a number of evenly distributed threshold points in the score range, at each of which accuracy, precision, and recall were assessed. These are presented in Table III.

TABLE III: Accuracy at binary image anomaly classification task, using a mean threshold on metric scores, and maximum accuracy achieved across the score range.

Method	**Accuracy (mean thresh)**	**Maximum Accuracy**
Residual	0.63	0.63
Discriminator scores (input)	0.52	0.56
Discriminator scores (regenerated)	0.56	0.59
PSNR	0.41	0.53
SSIM	0.41	0.52
AnoGAN-like score	0.63	0.63
Discriminator score difference	0.70	0.56

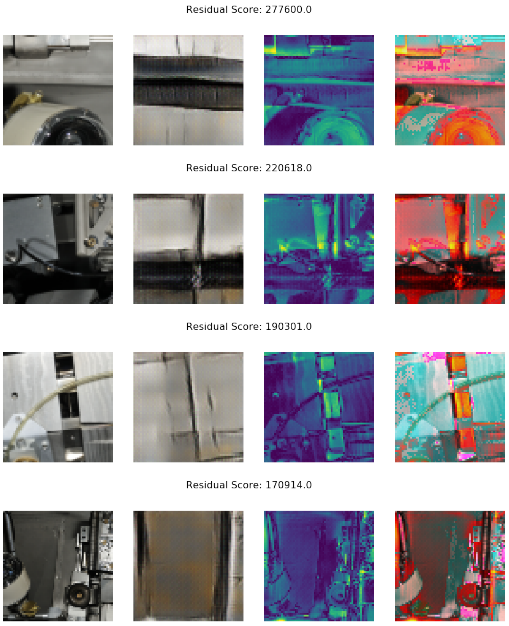

Fig. 6: Examples of anomalous query images and their regenerations (Columns 1 and 2 respectively), as well as residual images (Column 3) and detected anomalous regions (Column 4).

As can be clearly seen, none of the methods performed particularly well at this challenging task, although methods using the residual score outperform the others in terms of maximum accuracy.

A qualative assessment of the results, examples provided in Figures 6 and 7 gives some suggestions as to why the performance may not be ideal. Although the generative network is unable to accurately recreate the novel features, the method for predicting the latent representation using the full query image is taking the anomalous regions into account, and therefore the latent representation found geenrates an image which is not only visually close to the non-anomalous portions of the query image, but tries also to be close to the anomalous portions.

VI. Conclusions

We have demonstrated a new method for detecting anomalous images without any supervised training, and therefore without any prior knowledge of the nature of the anomalies to be detected. Regenerating images using GANs allows accurate reconstructions of images based on learned generative models, which represent the normal distribution of the image data, or external environment. The latent regressor method allows systems to regenerate images with an accuracy slightly lower than state-of-the-art methods, whilst improving speed of processing by several orders of magnitude, and thus bringing the techniques into applicability for real-time applications. In some cases, the method was able to regenerate details that were not recreated by the backpropagation method, although, in general both quality metrics still scored higher for the backpropagated regenerations. This perhaps indicates limitations in the PSNR and SSIM metrics for measuring reconstruction accuracy and visual similarity.

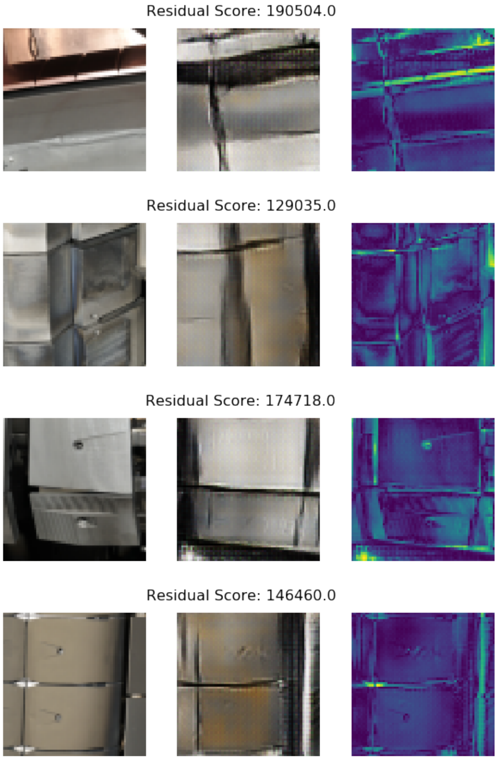

Fig. 7: Several examples of input images which do not contain anomalies (left), the corresponding regenerated images (middle), and residual images (right). The regenerations are fairly similar to the originals, and hence the residual images do not highlight anything of particular interest.

Image regeneration is of significant interest in novelty and anomaly detection, and the proposed methods can be further explored in relation to novelty-based saliency mapping, with application to robot interactions with an environment, and autonomous inspection. A number of possible improvements to the presented techniques provide further opportunities for valuable future work. These include combining the two herein compared methods, and using predicted latent representation as starting point for further iterative refinements.

In addition, we would like to explore applying newer architectures such as the BiGAN system for similar anomaly detection problems.

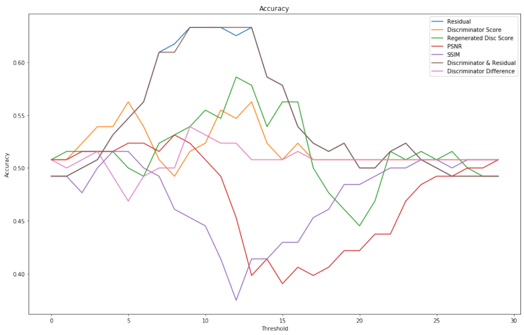

Fig. 8: Accuracy curves for each of the metrics.

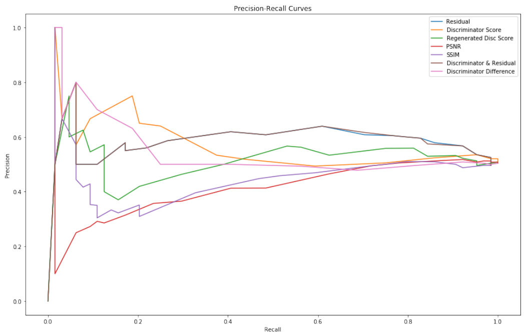

Fig. 9: Precision-Recall curves for each of the metrics.

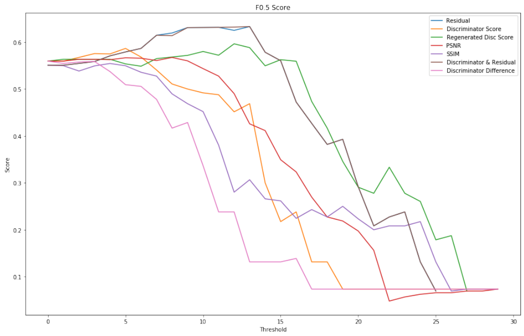

Fig. 10: $F_{0.5}$ scores for each of the metrics.

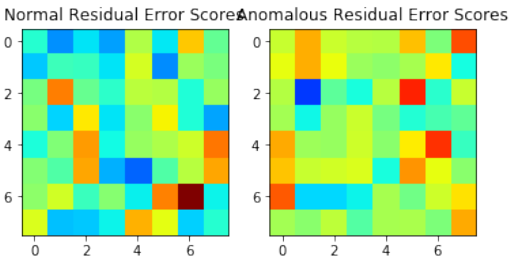

Fig. 11: Heat mapped illustrations of the residual anomaly scores of each of 64 test images from the non-anomalous set (left) and the anomalous set (right).

VII. Acknowledgements

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[5] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[6] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models.

[7] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[10] Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):331–345, 2014.

[11] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.