f-AnoGAN: Fast Unsupervised Anomaly Detection with Generative Adversarial Networks

Article *in* Medical Image Analysis · January 2019

DOI: 10.1016/j.media.2019.01.010

5 authors, including:

Thomas Schlegl
Medical University of Vienna
22 PUBLICATIONS **552** CITATIONS
SEE PROFILE

Philipp Seeböck
Medical University of Vienna
14 PUBLICATIONS **328** CITATIONS
SEE PROFILE

Sebastian M. Waldstein
Medical University of Vienna
71 PUBLICATIONS **1,172** CITATIONS
SEE PROFILE

Georg Langs
Medical University of Vienna
265 PUBLICATIONS **3,242** CITATIONS
SEE PROFILE

f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks

Thomas Schlegl[a,b], Philipp Seeböck[a,b], Sebastian M. Waldstein[b], Georg Langs[a,*], Ursula Schmidt-Erfurth[b]

[a] *Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria*
[b] *Christian Doppler Laboratory for Ophthalmic Image Analysis, Department of Ophthalmology and Optometry, Medical University Vienna, Austria*

Abstract

Obtaining expert labels in clinical imaging is difficult since exhaustive annotation is time-consuming. Furthermore, not all possibly relevant markers may be known and sufficiently well described a priori to even guide annotation. While supervised learning yields good results if expert labeled training data is available, the visual variability, and thus the vocabulary of findings, we can detect and exploit, is limited to the annotated lesions. Here, we present *fast AnoGAN* (*f-AnoGAN*), a generative adversarial network (GAN) based unsupervised learning approach capable of identifying anomalous images and image segments, that can serve as imaging biomarker candidates. We build a generative model of healthy training data, and propose and evaluate a fast mapping technique of new data to the GAN's latent space. The mapping is based on a trained encoder, and anomalies are detected via a combined anomaly score based on the building blocks of the trained model – comprising a discriminator feature residual error and an image reconstruction error. In the experiments on optical coherence tomography data, we compare the proposed method with alternative approaches, and provide comprehensive empirical evidence that *f-AnoGAN* outperforms alternative approaches and yields high anomaly detection accuracy. In addition, a visual Turing test with two retina experts showed that the generated images are indistinguishable from real normal retinal OCT images. The f-AnoGAN code is available at https://github.com/tSchlegl/f-AnoGAN.

Keywords: Anomaly detection, Wasserstein generative adversarial network, unsupervised learning, optical coherence tomography

1. Introduction

The detection and localization of imaging biomarkers correlating with disease status is important for initial diagnosis, assessment of treatment response and follow-up examinations. Spiculation patterns of lung nodules in lung CT scans (Zwirewich et al., 1991), microcalcification in X-ray mammography images for breast screening (Wang et al., 2014), or macular fluid in OCT scans of the retina (Schmidt-Erfurth et al., 2018) are examples of imaging biomarkers used in clinical routine. Training of highly accurate deep learning methods for the identification of imaging biomarkers has shown promising results reaching clinical expert level accuracies, but requires expert annotated data (Kooi et al., 2017; Esteva et al., 2017; Rajpurkar et al., 2017; Grewal et al., 2017). In practice, expert annotations suffer from two limitations. First, their number is typically limited due to the time costly acquisition, specifically for difficult to identify findings for which machine learning approaches would be particularly desirable. Second, even if annotated training corpora are available, supervised learning is limited to already known markers. In some contexts, they exhibit high inter rater variability and correspondingly limited prediction power (Walsh et al., 2015), and we suspect that relevant markers exist beyond those already

*Corresponding author: Georg Langs , Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab, Medical University Vienna, Austria, www.cir.meduniwien.ac.at
Email addresses: thomas.schlegl@gmail.com (Thomas Schlegl), georg.langs@meduniwien.ac.at (Georg Langs)

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

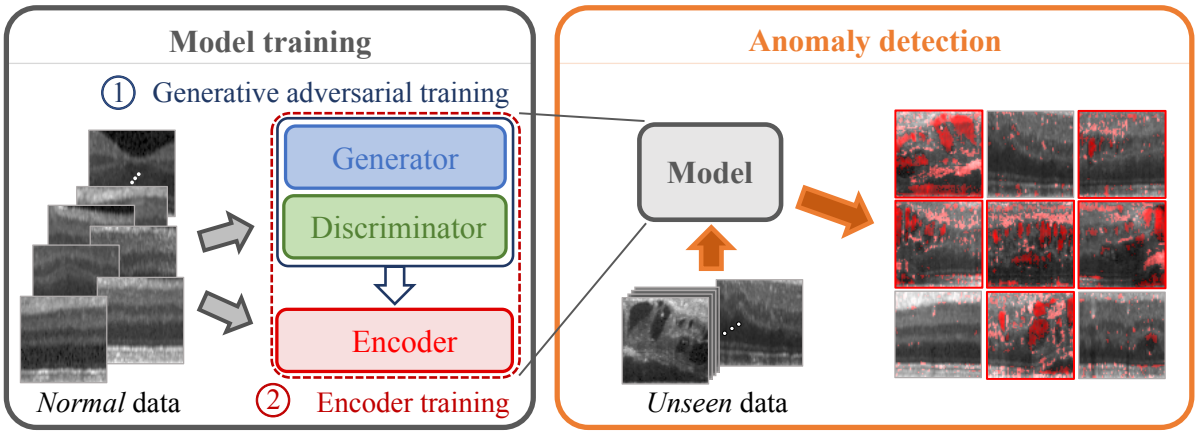

Figure 1: Anomaly detection framework. Both steps of model training, generative adversarial training (yields a trained generator and discriminator) and encoder training (yields a trained encoder), are performed on *normal* ("healthy") data and anomaly detection is performed on both, unseen healthy cases and anomalous data. (Best viewed in color)

described. Here, we propose a fast anomaly detection technique trained on large-scale imaging data only comprising normal images without the need of annotations as learning targets during training. Only for data selection prior to training, volume-level information is needed, namely to select imaging data for training that show solely normal appearance. We perform unsupervised learning on these data to train a generative model that captures a high amount of natural (normal) variability inherent in the data used during training. Subsequently, we train an encoder to enable fast mapping of images to latent space [1] and thus facilitate a fast evaluation of whether or not novel images and image segments fit the manifold of normal data (Figure 1).

Capability and limitations of supervised learning. Automated classification and segmentation of anatomical structures, lesions and imaging patterns has been heavily studied for the last decades in medical image analysis (Esteva et al., 2017; Schlegl et al., 2017b; Ramteke and Monali, 2012; Weisenfeld et al., 2006; Anbeek et al., 2004; Kaus et al., 2001; Van Leemput et al., 1999; Reddick et al., 1997; Stansfield, 1986). Deep learning provides powerful methods for automated analysis of large clinical data sets including electronic health records (Miotto et al., 2016) or radiology reports (Chen et al., 2017; Schlegl et al., 2015). It has been applied to medical images, such as for example X-ray (Rajpurkar et al., 2017; Kooi et al., 2017), computed tomography (CT) (Grewal et al., 2017), magnetic resonance imaging (MRI)(Kamnitsas et al., 2017), optical coherence tomography (OCT) (Schlegl et al., 2017b), or histology data (Cireşan et al., 2013)), or high-throughput sequencing data (Alipanahi et al., 2015), enabling reproducible and higher quality accurate decisions. Learning a discriminative model requires supervision in the form of annotations (e.g. positive (anomalous) vs. negative (normal) class) during model training. This comes with two drawbacks. First, annotations are time-consuming and require thorough domain knowledge. Secondly, the vocabulary of detectable anomalies by a trained model is restricted to those types already known and defined as targets for training. In contrast, unsupervised learning can model the distribution of normal observations, and enables the identification of observations – *anomalies* – that do not fit this model well. Anomaly detection is useful in clinical routine, to screen large amounts of data for potentially relevant findings, that should be further assessed by experts. Furthermore, anomaly detection enables the expansion of the vocabulary of markers used for quantitative disease and treatment assessment. In both cases, the identification of anomalous images, and the localization of anomalies on the pixel level are relevant. *f-AnoGAN* allows for anomaly detection on the image level and localization of anomalies on the pixel level.

[1] In the context of generative adversarial networks, the *latent space* is also termed *z-space*. We use both terms interchangeably.

Anomaly detection. Anomaly detection is the identification of data that does not fit to the distribution of normal data, i.e. does not conform to normal appearance, semantic content, quality, or expected behavior. Anomaly detection techniques are employed in various domains, ranging from lesion detection in medical imaging (Schlegl et al., 2017a), to credit-card fraud detection in finance (Awoyemi et al., 2017), to radio frequency anomaly detection in wireless networks (Tandiya et al., 2018). Novelty detection (Sabokrou et al., 2018; Pimentel et al., 2014) is the identification of new, yet unobserved patterns in test data not included in training data. Since anomaly detection and novelty detection share the same methodological foundation, the proposed anomaly detection framework can also be applied for novelty detection. The main focus of this paper is to propose methodology and to evaluate its capability to detect (known) anomalies not present in the training data. We leave the distinction and the evaluation of the capacity of detected anomalies as (potentially novel) biomarkers to future studies. Anomaly detection techniques comprise statistical approaches (Nguyen and Goulet, 2017), density-based anomaly detection (Zhang et al., 2018), clustering-based anomaly detection (Alguliyev et al., 2017), graph-based anomaly detection (Akoglu et al., 2015), or support vector machine (SVM)-based anomaly detection (Erfani et al., 2016; Seeböck et al., 2018), to name but a few examples.

Related work. The proposed algorithm is based on generative adversarial networks (GANs), introduced by Goodfellow et al. (2014) and capable of generating realistic images (Radford et al., 2015). A GAN consists of two adversarial networks, a *generator* G and a *discriminator* D. In the first formulation of GAN training (Goodfellow et al., 2014), the loss function quantified the Jensen-Shannon (JS) divergence between the training data distribution $\mathbb{P}_r$ and the generator sample distribution $\mathbb{P}_g$ defined by $\tilde{\mathbf{x}} = G(\mathbf{z})$, with $\mathbf{z} \sim p(\mathbf{z})$. This formulation suffers from training instability and is prone to mode collapse, and thus GANs were hard to train. Arjovsky et al. (2017) proposed to replace JS divergence by the Wasserstein distance since the smoother value space stabilizes the training procedure. In the Wasserstein GAN (WGAN) formulation, the discriminator does not directly discriminate between real and generated samples but estimates the Wasserstein distance between the generator and the real data distribution. For enforcing a Lipschitz constraint, after every gradient update, the discriminator weights are clipped to a small range $[-c, c]$ bounded by constant hyperparameter c. Since this is a "terrible way to enforce a Lipschitz constraint" (Arjovsky et al., 2017), Gulrajani et al. (2017) proposed an improved WGAN training procedure replacing weight clipping by *gradient penalty*, where the gradient norm of the discriminator's output is directly constrained with respect to the discriminator's input.

The presented work is most closely related to our conference paper proposing *AnoGAN* (Schlegl et al., 2017a), a GAN based anomaly detection technique. We utilized a deep convolutional generative adversarial network (DCGAN) (Radford et al., 2015) to train a generator and discriminator on normal data via unsupervised learning. During detection, the latent space location for a given query image was determined performing iterative backpropagation leading to an anomaly score under utilization of the trained generator and discriminator. Both approaches, *AnoGAN* as well as the presented *f-AnoGAN*, require the mapping from image to latent space. The recently proposed *inversion* method of Creswell and Bharath (2018) is similar in spirit to the iterative mapping approach from images to latent vectors. Lipton and Tripathi (2017) proposed *stochastic clipping* for an improved latent vector recovery using an iterative mapping scheme as well. Although *AnoGAN* showed good anomaly detection performance, iterative approaches share the drawback on real world applications of computational inefficiency during detection time. The *f-AnoGAN* technique replaces this iterative procedure by a learned mapping from image to latent space, dramatically improving speed.

In the *ALI* (*Adversarially Learned Inference*) (Dumoulin et al., 2016) and BiGAN (*Bidirectional Generative Adversarial Network*) (Donahue et al., 2016) models, the mapping between image space and latent space is jointly learned in both directions during GAN training. Donahue et al. (2016) trained a *"latent regressor"* to map images to latent space and used it as baseline for the BiGAN model. Although architecturally similar to the *ziz architecture* (Section 2.2.1), one of the examined encoder training approaches, they didn't study the performance of *latent regressor* encoder training in an anomaly detection setting. An *adversarial convolutional autoencoder (AdvAE)* also jointly trains an encoder and a decoder based on a loss function that combines an image residual term and an adversarial term. In literature, we find utilization of *AdvAEs* for example for image inpainting (Pathak et al., 2016).

Since the publication of the *AnoGAN* technique, further applications of GAN-based anomaly detection followed. Zheng et al. (2018) proposed a GAN-based telecom fraud detection approach, where a deep denoising autoencoder (AE) learns the relationship among the inputs and adversarial training is employed to discriminate between positive and negative samples in the data distribution. Ravanbakhsh et al. (2017) employed two conditional generators, generating an optical flow image conditioned by a video frame and vice versa. The discriminator takes as input two images and decides whether both images are real data samples. The training procedure followed our proposed idea to the effect that only normal data was used for training. Their approach does not include a separate encoder training procedure as it is used in the present work. Following *AnoGAN*, Zenati et al. (2018) trained a GAN-based architecture on normal samples only and utilized the *AnoGAN* anomaly scoring function. To speed up anomaly scoring during test time, they simultaneously trained the encoder during GAN training by employing the *BiGAN* model, which is methodically equivalent to the *ALI* model that is utilized as baseline approach in our experiments.

Anomaly detection on OCT images was also studied by Seeböck et al. (2018) employing a convolutional AE and a one-class SVM. A drawback of this approach is the need for choosing a prior on the healthy amount of the volume. Sidibe et al. (2017) performed anomaly detection on OCT scans and modeled normal appearance of OCT images with a Gaussian Mixture Model (GMM) trained on a PCA embedding of OCT images. Thus this approach enables the detection of anomalous images but the localization of anomalous regions within an image is not possible. Besides those unsupervised learning approaches, we find classifier training approaches on OCT data, namely supervised learning (Venhuizen et al., 2015; Schlegl et al., 2017b), semi-supervised learning (Farsiu et al., 2008), or weakly supervised learning (Schlegl et al., 2015), that all require some form of supervision signal (i.e. annotations as target labels) and thus are restricted to a predefined set of lesions defined by the class labels.

Contribution. We propose fast anomaly detection based on GAN training on normal image appearance (see black block in Figure 2) capturing normal variability of training data, described in Section 2.1, and a subsequent encoder training approach (see red blocks in Figure 2), enabling fast mapping from images to corresponding locations of the learned latent representation (Section 2.2) in a single step during inference. This makes the proposed technique suitable for real-time anomaly detection applications. We present an in-depth study of different encoder training approaches, the *"latent regressor"* architecture, a z-to-z mapping (*ziz*) procedure known from literature (Donahue et al., 2016), and beyond that two new image-to-image (*izi*) mapping approaches. Encoder training for mapping images to latent vectors needs a pre-trained GAN. We use a WGAN, a state-of-the-art GAN, architecture. However, our proposed approach is not limited to WGAN training but can be applied to any pre-trained GAN. To the best of our knowledge, in this work, a discriminator guided image-to-image mapping approach (izi_f) is utilized for the first time for learning the mapping from images to latent encodings in a subsequent training step. In comparison to the investigated alternative approaches, this novel approach yields the best anomaly detection as well as the best anomaly localization (i.e. segmentation) performance. The *f-AnoGAN* technique is motivated by our initial work on anomaly detection, where a similar approach is used in an iterative manner. Both, the application of WGAN training to learn a representation of normal image variability and the application of the examined encoder training approaches to anomaly detection is new in general.

The present work differs in several aspects from the *AnoGAN* paper (Schlegl et al., 2017a): 1) it uses the state-of-the-art improved WGAN architecture instead of the DCGAN architecture, 2) it introduces an approach to substantially speed up the mapping of input images to the latent space during the detection by moving from an iterative gradient descent approach to a learned mapping. Finally, it expands the experimental evaluation of the algorithm to a wide range of alternatives.

2. Fast GAN based anomaly detection

The proposed anomaly detection framework consists of two training steps on normal images: (1) GAN training, and (2) encoder training based on the trained GAN model. After training, inference yields an anomaly score for a new image utilizing these trained components. Similar to Schlegl et al. (2017a), we train a generative adversarial network on images from normal scans, yielding a generator

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

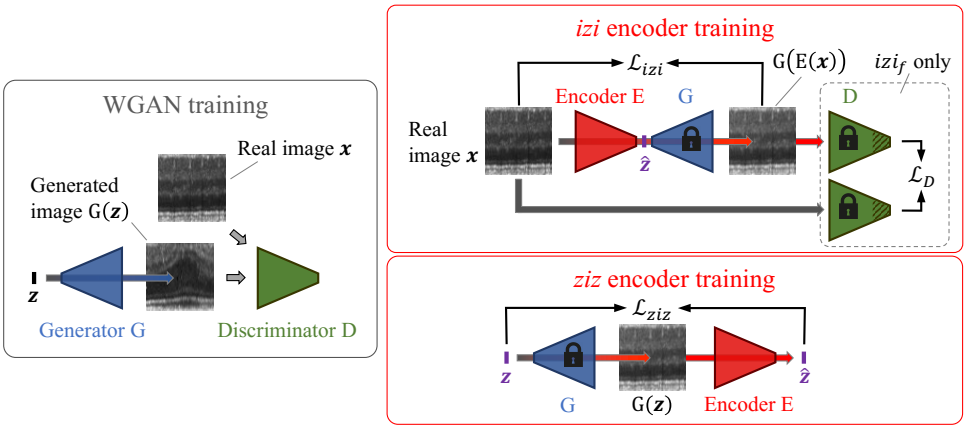

Figure 2: Components of the proposed anomaly detection framework training. *Wasserstein GAN (WGAN)* training yielding learned parameters for the generator (G) and discriminator (D). Three possible variants of *encoder training* with fixed parameters of G (and D) where only the encoder parameters are adapted. 1) *izi encoder training*: minimizing the loss $\mathcal{L}_{izi}$ based on the residual of real input images and *"reconstructed"* images, 2) izi_f *encoder training*: jointly minimizing the loss $\mathcal{L}_{izi}$ based on the residual of real input images and *"reconstructed"* images and the loss $\mathcal{L}_{\mathcal{D}}$ based on the residual on discriminator's features, and 3) *ziz encoder training*: minimizing the loss $\mathcal{L}_{ziz}$ based on the residual of randomly sampled and reconstructed locations in z-space (latent space).

and discriminator, and a latent representation of normal anatomical variability. Using this model, we train an encoder that maps images to the latent space (Figure 2). After encoder training, the encoder maps images to locations in the latent space that map to the normal version of the input image, when fed as input to the trained generator. In the case of a normal input image (and under the assumption of a perfect generator and a perfect encoder), mapping from image space to the latent space via the encoder and subsequent mapping from latent space back to image space via the generator should closely resemble the identity transform. The degree of deviation is used for anomaly scoring. Both components of the anomaly detection framework training are illustrated in Figure 2.

In the following, we give a description of generator and discriminator training (Section 2.1), followed by a detailed description of encoder training (Section 2.2), and anomaly scoring used for the identification of anomalous appearance (Section 2.3).

2.1. Unsupervised learning of normal anatomical variability

For the training of the GAN and the encoder, we use unlabeled training data $\mathbf{x} = \mathbf{x}_{k,n} \in \mathcal{X}$ sampled from the manifold $\mathcal{X}$ of images showing normal anatomy, where $\mathbf{x}_{k,n}$ are image patches of size $s \times s$ extracted at K randomly sampled positions from image $\mathbf{I}_n$, with $k = 1, 2, \ldots, K$. We use a set of N medical images $\mathbf{I}_n$ that show normal anatomy only, with $n = 1, 2, \ldots, N$, where $\mathbf{I}_n \in \mathbb{R}^{u \times v}$ is an intensity image of size $u \times v$, with $u \gg s$ and $v \gg s$. For evaluation, we use image data $\mathbf{y}_m$ of size $s \times s$ extracted from testing data $\mathbf{J}$ unseen during training and corresponding pixel-level binary annotations $\mathbf{a}_m \in \{0, 1\}^{s \times s}$ of the same size, with $m = 1, 2, \ldots, M$. This test set $\langle \mathbf{y}_m, \mathbf{a}_m \rangle$ comprises tuples of both, normal and anomalous images, and corresponding labels, and is only used for the purpose of pixel-level anomaly *localization* evaluation. For quantitative anomaly *detection* performance evaluation, we use image-level labels $\dot{a}_m \in \{0, 1\}$. During training, only unlabeled normal images are used.

Learning normal anatomical variability with a generative adversarial network. We train a WGAN to learn a non-linear mapping function from latent space $\mathcal{Z}$ to the manifold $\mathcal{X}$ in the image space that represents the variability of (normal) training images only based on the set of unlabeled images $\langle \mathbf{I}_n \rangle$. During GAN training, the generator G and the discriminator D are simultaneously optimized. Given input noise sampled from latent space $\mathcal{Z} \in \mathbb{R}^d$ of dimension d, the generator is trained to output samples that follow the distribution $\mathbb{P}_g$ over data $\mathcal{X}$ as close as possible to the distribution of real data $\mathbb{P}_r$, and thus to fool the discriminator. Here, the generator learns to generate images of the training distribution capturing normal variability. The discriminator can estimate the fit of generated images to the distribution of training images. The trained generator and discriminator are utilized with fixed weights for subsequent encoder training (Section 2.2) and for anomaly scoring (Section 2.3).

2.2. Learning a fast mapping from images to encodings in the latent space

GAN training yields a generator $G(\mathbf{z}) = \mathbf{z} \mapsto \mathbf{x}$ that maps from $\mathcal{Z}$ to $\mathcal{X}$, but not the inverse mapping from $\mathcal{X}$ to $\mathcal{Z}$ that is needed in our anomaly detection technique. We learn the mapping $E(\mathbf{x}) = \mathbf{x} \mapsto \mathbf{z}$ by training a deep encoder network E. The encoder can be trained with either of two basic architectures: (1) *z-image-z* (in the following abbreviated as *ziz*) encoder training, or (2) *image-z-image* (in the following abbreviated as *izi*) encoder training. In both cases we use a convolutional autoencoder (AE) architecture, comprising a trainable encoder E that maps from image to z-space and the generator that acts as decoder by mapping from z to image space with fixed weights resulting from WGAN training. Both differ in the order of utilization of the encoder and the decoder (i.e. trained generator). During encoder training, only the encoder parameters are optimized, whereas the generator parameters are kept fixed. The investigated encoder training architectures are illustrated in Figure 2 (red blocks).

2.2.1. Training the encoder with generated images: ziz architecture

Reversing the order regarding encoder and decoder (i.e. the pre-trained generator in our specific case) utilization in a standard AE leads to the **ziz architecture**. During training, a random sample drawn from z-space is mapped to the image space through the fixed generator G, and the encoder E is trained to map it back to the z-space. Therefore, for *ziz* encoder training, no imaging data is needed. The *ziz* architecture resembles a z-to-z AE, where the z-to-image mapping G is fixed. During training, we minimize the mean squared error (MSE) of input z-samples $\mathbf{z}$ and reconstructed z-samples $E(G(\mathbf{z}))$:

$$\mathcal{L}_{ziz}(\mathbf{z}) = \frac{1}{d}\|\mathbf{z} - E\big(G(\mathbf{z})\big)\|^2, \tag{1}$$

where d is the dimensionality of the z-space.

This approach is similar to the "latent regressor" that was used as baseline model in a DCGAN architecture for the BiGAN model in the *"Adversarial feature learning"* work of Donahue et al. (2016). In contrast to the *izi architecture*, here we know the true target z location. The drawback of this approach is that, in contrast to the *izi architecture*, the encoder only "sees" generated images but never receives real input images (Donahue et al., 2016).

2.2.2. Training the encoder with real images: izi architecture

The **izi architecture** follows a standard AE configuration, where an encoder is followed by a decoder (generator). During training, the mapping from real images to latent encodings $\mathbf{z}$ is performed by the trainable encoder, while the mapping of $\mathbf{z}$ back to image space is performed with the fixed generator G. This architecture resembles an image-to-image AE. We minimize the MSE residual loss of input images $\mathbf{x}$ and reconstructed images $G(E(\mathbf{x}))$:

$$\mathcal{L}_{izi}(\mathbf{x}) = \frac{1}{n}\|\mathbf{x} - G\big(E(\mathbf{x})\big)\|^2, \tag{2}$$

where $\|\cdot\|^2$ is the sum of squared pixel-wise residuals of gray values and n is the number of pixels in an image. The *izi* encoder is trained with the same data used for WGAN training, i.e. only normal images. This approach has an important drawback. Since the *true* target location in the z-space of a given query image is unknown, we can only indirectly measure the accuracy of the image to z mapping through mapping back to the image space and computing the image-to-image residual.

2.2.3. Discriminator guided izi encoder training: izi_f architecture

The *izi* training objective enforces similarity in the image space. The mapping of new images can result in positions in regions of the latent space only sparsely sampled during training that would not convince the discriminator, when mapped back to image space. As a consequence, only minimizing pixel-wise differences will occasionally yield images that are not realistic examples of normal images, but still have a small residual even for anomalous images. This makes the residual in the image space not a reliable solitary marker of anomaly.

We found that the residual in the feature space populated by the discriminator, which is a reliable basis for identifying anomalous images, is an essential term in the encoder training objective. Therefore,

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

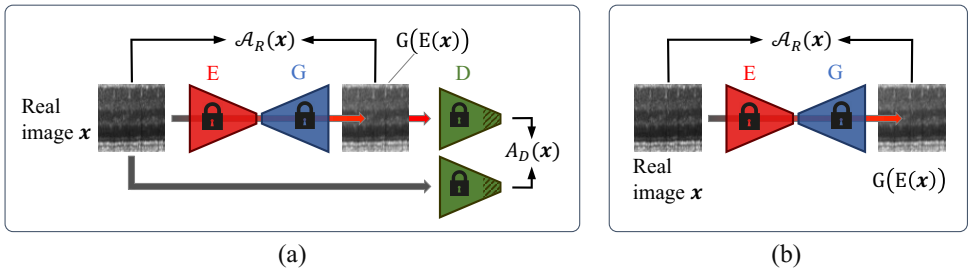

Figure 3: Inference on new images for anomaly scoring. Anomaly scoring uses exactly the same architecture and dataflow used for encoder training. (a) Proposed *f-AnoGAN* model that uses discriminator guided encoder training (izi_f architecture). (b) Anomaly quantification for both underlying encoder training architectures that do not include a discriminator based term (*izi* architecture and *ziz* architecture). (Best viewed in color)

we additionally calculate image statistics of the real image and the reconstructed image resulting in the **izi$_f$ architecture**. The loss function for discriminator guided *izi* encoder training (in the following abbreviated as izi_f) is:

$$\mathcal{L}_{izi_f}(\mathbf{x}) = \frac{1}{n} \cdot \|\mathbf{x} - G\big(E(\mathbf{x})\big)\|^2 + \frac{\kappa}{n_d} \cdot \|f(\mathbf{x}) - f\big(G\big(E(\mathbf{x})\big)\big)\|^2, \tag{3}$$

where discriminator features $f(\cdot)$ of an intermediate layer are used as statistics of a given input, n_d is the dimensionality of the intermediate feature representation, and κ is a weighting factor. Using discriminator features is inspired by the feature matching technique proposed in Salimans et al. (2016), and is also related to the loss used for iterative z-mapping in our initial anomaly detection work (Schlegl et al., 2017a). The parameters of the discriminator are those learned during WGAN training and are kept fixed during encoder training. Since the **izi$_f$ architecture** simultaneously guides encoder training in the image space and in the latent space, we choose izi_f as the proposed encoder training architecture in the *f-AnoGAN* framework.

2.3. Detection of anomalies

During image-level anomaly detection, we quantify the deviation of query images and corresponding "reconstructions". All the components, needed for generating the image reconstructions and performing the anomaly quantification are trained in the WGAN training step (Section 2.1) and in the encoder training step (Section 2.2).

The anomaly quantification formulation follows directly the specific definition of the loss used for encoder training (Equations (1) to (3)). For the proposed *f-AnoGAN* model, which implements the discriminator guided izi_f encoder training (Equation (3)), the final anomaly score $\mathcal{A}(\mathbf{x})$ for a new image $\mathbf{x}$ is defined by

$$\mathcal{A}(\mathbf{x}) = \mathcal{A}_R(\mathbf{x}) + \kappa \cdot \mathcal{A}_D(\mathbf{x}), \tag{4}$$

where $\mathcal{A}_R(\mathbf{x}) = \frac{1}{n} \cdot \|\mathbf{x} - G\big(E(\mathbf{x})\big)\|^2$, $\mathcal{A}_D(\mathbf{x}) = \frac{1}{n_d} \cdot \|f(\mathbf{x}) - f\big(G\big(E(\mathbf{x})\big)\big)\|^2$ and κ again is a weighting factor (Figure 3a). The definition of the anomaly score for both encoder training architectures that do not include a discriminator based term, *izi* architecture (Equation (2)) and *ziz* architecture (Equation (1)), reduces to $\mathcal{A}(\mathbf{x}) = \mathcal{A}_R(\mathbf{x}) = \frac{1}{n} \cdot \|\mathbf{x} - G\big(E(\mathbf{x})\big)\|^2$ (Figure 3b).

In general, both formulations yield high anomaly scores on anomalous images and small anomaly scores on normal input images. Since the model is only trained on normal images, it only "reconstructs" an image visually similar to the input image and lying on the manifold of normal images $\mathcal{X}$. The ability of reconstructing visually similar images is inversely proportional to the degree or distinction of anomaly. Normal query images result in small deviations whereas anomalous images are mapped to "reconstructions" yielding large deviations. The absolute value of pixel-wise residuals $\dot{\mathcal{A}}_R(\mathbf{x})$, defined by:

$$\dot{\mathcal{A}}_R(\mathbf{x}) = |\mathbf{x} - G\big(E(\mathbf{x})\big)|, \tag{5}$$

is used for pixel-level anomaly *localization*.

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

3. Experiments

We evaluate the proposed anomaly detection technique on optical coherence tomography imaging data of the retina. We examine if the model can generate realistic images, if it can identify images containing anomalies, and localize the anomalous regions. We compare anomaly scoring accuracy of our proposed approach with the following alternative approaches: convolutional autoencoder (*AE*), adversarial convolutional autoencoder (*AdvAE*), *ALI* model. Furthermore, we compare *f-AnoGAN* with an *iterative* backpropagation based approach on a trained WGAN following Schlegl et al. (2017a). Additionally, we evaluate the performance of an alternative anomaly scoring approach, A_D, that only utilizes the discriminator output $D(\mathbf{x})$ of the trained WGAN. In contrast to DCGAN training, where the discriminator outputs a measure of "realness" of a given input image, in the WGAN formulation the discriminator estimates the Wasserstein distance between the generator and the real data distribution. Hence the discriminator's output can not be directly used for anomaly scoring. To obtain the anomaly score for an unseen testing image $\mathbf{x}_u$ based on the WGAN value function, we compute the distance:

$$A_D = \hat{m}_{x_t} - D(\mathbf{x}_u) \tag{6}$$

between the discriminator output $D(\mathbf{x}_u)$ on the given query image $\mathbf{x}_u$ and the average discriminator output

$$\hat{m}_{x_t} = \mathop{\mathbb{E}}_{\mathbf{x}_t \sim \mathbb{P}_r} (D(\mathbf{x}_t)) \tag{7}$$

over 32.000 randomly chosen training images $\mathbf{x}_t$.

Data, data selection and preprocessing. We used the same clinical high resolution retinal spectral-domain optical coherence tomography (SD-OCT) volumes (acquired with a Heidelberg Spectralis device) as described in Schlegl et al. (2017a) to facilitate comparability. Each SD-OCT volume comprises 49 cross-sectional images (acquired in zx-direction). The total volume resolutions in z-, x-, and y direction are $496 \times 512 \times 49$ voxels, respectively. The following preprocessing steps were applied to every SD-OCT volume. The voxel gray values were normalized per volume to range from -1 to 1. The number of columns was reduced from 512 to 256, resulting in a pixel size (in x-direction) of $22\mu m$. To counterbalance variations in shape, thickness and spatial orientation of the retina among individual OCT images and patients, we extracted and flattened the retinal area in the following way. The area of the retina was localized via an automatic layer segmentation algorithm (Garvin et al., 2009), yielding for every column the locations of the top and bottom retinal layer. We used the bottom layer as fixed boundary, and only if the retinal thickness (defined as the number of pixels between top and bottom layer) exceeded the extracted area height, we normalized thickness, if it was less we did not alter it.

WGAN and subsequent encoder training was performed on approximately 850,000 2D image patches with 64×64 pixels extracted at randomly sampled positions from 270 SD-OCT volumes of healthy subjects for whom the absence of retinal fluid was confirmed. Quantitative and qualitative evaluation was performed on a separate test set – comprising 10 SD-OCT volumes of healthy subjects and 10 diseased cases containing retinal fluid – not used during WGAN and encoder training. For these volumes we also had pixel-level binary annotations (fluid vs. non-fluid) provided by clinical retinal experts. These annotations were only used for evaluation of the anomaly localization performance but were not used during training. Quantitative image-level anomaly detection accuracy was evaluated based on the occurrence of at least a single pixel annotated as retinal fluid in the image patch. The test set was composed of normal and pathological samples and in total consisted of 70,000 2D image patches. The amount of normal and anomalous (i.e. showing retinal fluid) image patches in the test set was 70% and 30% respectively. Figure 4 shows an overview of the data preprocessing steps from OCT scans to 2D image patches.

3.1. Evaluation

For evaluation of the WGAN training, encoder training, anomaly detection and anomaly localization performance we performed the following experiments.

(1) Does the model capture normal variability? The quantification of the quality of generative models is an open question. Salimans et al. (2016) proposed the *Inception Score*, the most widely used

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

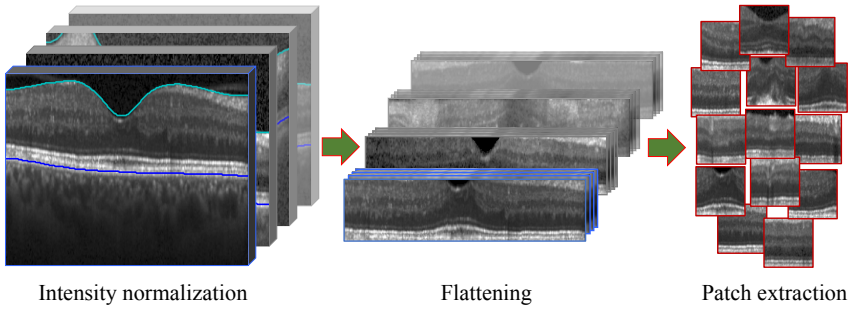

Figure 4: Data preprocessing. 1) Volume-wise intensity normalization, 2) extraction and flattening of the retinal area, 3) 2D patch extraction. (Best viewed in color)

metric for evaluating image-generating models. Barratt and Sharma (2018) discovered several shortcomings within the underlying premise of the score and its application, as for instance both slight changes in network weights and the specific Inception network implementation result in different Inception Scores for the same set of generated images. Moreover, the original Inception Score is applicable only for ImageNet generators, and application of the Inception Score on other datasets requires training of an Inception network on the specific dataset in question (cf. (Barratt and Sharma, 2018)), which limits the comparability of the Inception Score across medical imaging datasets. Therefore, we presented a visual Turing test in the spirit of Chuquicusma et al. (2018) to two clinical retina experts in order to quantify the quality of the generated normal images. For the evaluation of the expert's perception of both real and generated images, we used a set of 100 images with 64×64 pixels that comprised 50 real normal retinal OCT images and 50 generated images, which were generated using only the trained generator from WGAN training conditioned by randomly sampled locations in the latent space. Both clinical retina experts independently rated the realness of the images and were blinded of each other's scoring.

Furthermore, we explored qualitatively whether the model is able to capture the visible variability of real normal images. Furthermore, we analyzed whether the learned latent representation is *smooth*. Unstable GAN training can lead to mode collapse, manifesting in a state of only few locations in z-space (latent space), which allow the generator to generate realistic images. We tested for smooth coverage of the latent space, by choosing random z positions, and generating image series from sampling points along the connecting line. Smooth transitions in image series indicate the absence of mode collapse. In an additional experiment, we conditioned the start and end points of the sampling in z-space with real normal images, which were randomly selected from the training set.

(2) Does the model detect anomalous images accurately? Quantitative evaluation of the anomaly detection accuracy was performed on the annotated test data comprising normal and anomalous images. We compared the accuracy of ***f-AnoGAN*** with ***AE***, ***AdvAE***, and ***ALI*** model (Dumoulin et al., 2016), which is methodically equivalent to the *BiGAN* model (Donahue et al., 2016). **AE**, **AdvAE**, and **ALI** are valid alternatives to learn a mapping from images to latent encodings and back to image space. They jointly learn both mappings in one training phase. An *AE* is a basic model for joint encoder-decoder training, while *AdvAE* and *ALI* additionally incorporate adversarial training. But among the investigated alternative approaches, only a *BiGAN* based model was hitherto used for anomaly detection (Zenati et al., 2018). To maximize comparability, we trained the AE based models with the same network architectures as used in the WGAN generator and discriminator. More specifically, we implemented a *ResNet* (He et al., 2016) based architecture used in *f-AnoGAN* and Gulrajani et al. (2017) also in *AE* and *AdvAE*. The *ALI* model was trained in the originally proposed style. These approaches replaced both *f-AnoGAN* training procedures, WGAN and subsequent encoder training, while leaving as far as applicable the computation of the final anomaly score unchanged. Since a standard *AE* architecture does not comprise a discriminator, anomaly scoring was based on the image residual only; for *AdvAE* and *ALI* we added the discriminator feature residual analogously to *f-AnoGAN*. We emphasize that these approaches are valid alternatives implemented with the aim of being as competitive and as comparable as possible to *f-AnoGAN*.

Furthermore, we compared *f-AnoGAN* with two WGAN based alternative approaches that do not train a separate encoder: 1) using only the WGAN discriminator in a single feed-forward step ($\mathbf{A_D}$;

Equation (6)), and 2) using the trained WGAN generator and discriminator via iterative z-mapping during detection following Schlegl et al. (2017a) (***iterative***).[2] We show receiver operating characteristic (ROC) curves, and report corresponding area under the ROC curve (AUC) values, as well as precision, sensitivity, specificity, and f-score at the Youden index of the ROC curves on image-level anomaly detection performance for the proposed *f-AnoGAN* and all alternative approaches.
(3) Can the model localize anomalies in images? To assess the ability of anomaly detection to localize anomalies on pixel level, we show qualitative results comparing the residual difference between reconstructed images and real input images with ground-truth annotations of retinal fluid. In addition, quantitative evaluation of the pixel-level anomaly localization accuracy was performed on the annotated test data. We compared the localization accuracy of ***f-AnoGAN*** with ***AE***, ***AdvAE***, and ***ALI*** model.
(4) How does accuracy achieved with different encoder training architectures compare? We compared the final image-level anomaly detection accuracy of three different encoder training approaches (Figure 2): (1) *ziz encoder training*, (2) *izi encoder training*, or izi_f *encoder training* (i.e. the *f-AnoGAN* model). During inference on new images for anomaly scoring, all approaches use the same generator trained during WGAN training (Figure 3); only the izi_f additionally utilizes the trained discriminator (again obtained during WGAN training). Additionally, we performed qualitative evaluation of the ability to reconstruct query images and to localize anomalies.

Implementation details. We used the improved WGAN training procedure (Gulrajani et al., 2017) for stable GAN training. Following Gulrajani et al. (2017), we used a normally distributed z-space with 128 dimensions. We trained a ResNet based WGAN with gradient penalty (WGAN-GP) proposed by Gulrajani et al. (2017), where the generator and the discriminator were implemented as simple convolutional decoder and encoder respectively, each comprising four residual blocks.[3] From the first to the last residual block, the generator utilized $512 - 256 - 128 - 64$ and the discriminator utilized $128 - 256 - 512 - 512$ filter kernels. Throughout, filters of size 3×3 were used. Following Gulrajani et al. (2017), in the discriminator any batch normalization was replaced with layer normalization. We additionally implemented an encoder with an architecture similar to the generator's architecture (Section 2.2). WGAN training was performed as much as 100,000 iterations with batch size 64 that are 7 full utilizations of all training images, followed by 50,000 encoder training iterations with learned generator and discriminator parameters kept fixed. For WGAN training we used Adam (Kingma and Ba, 2014), a stochastic optimizer, whereas for subsequent encoder training *RMSprop* (Hinton et al., 2013), a widely used adaptive learning rate method, was utilized. As suggested by Gulrajani et al. (2017), we trained the discriminator more often than the generator. More specifically, after every generator update we ran 5 discriminator updates. For simplicity, we used $\kappa = 1.0$ in Equations (3) and (4), i.e. we weighted equally the image residual and discriminator feature residual terms in both, the encoder training loss function and the anomaly scoring function.

Sampling from a normal distribution during WGAN training provides the possibility to restrict the mapping to the part of the latent space that describes the main components of normal variability of training data whereas areas of the latent space are blocked out that were only sparsely sampled during training. Therefore, during encoder training, we restricted the encoder mapping to the range $(-1\sigma, +1\sigma)$ of the standard normal distribution by applying a *tanh* activation function on the encoder outputs. A more detailed motivation on constraining the encoder outputs can be found in Appendix A.2.

All experiments were performed on a Titan X graphics processing unit based on a system running Python 2.7, TensorFlow (Abadi et al., 2015) library (version 1.2) and CUDA 8.0. We provide code for training and inference of the full model. [4]

[2] During iterative z-mapping, we performed stochastic clipping in the range $[-1, 1]$ for an improved latent vector recovery as recently proposed by Lipton and Tripathi (2017).

[3] For generator- and discriminator training we adapted the Tensorflow-based implementation: `https://github.com/igul222/improved_wgan_training`.

[4] Code is provided at `https://github.com/tSchlegl/f-AnoGAN`, and together with detailed information at `http://www.cir.meduniwien.ac.at/research/anomalies`.

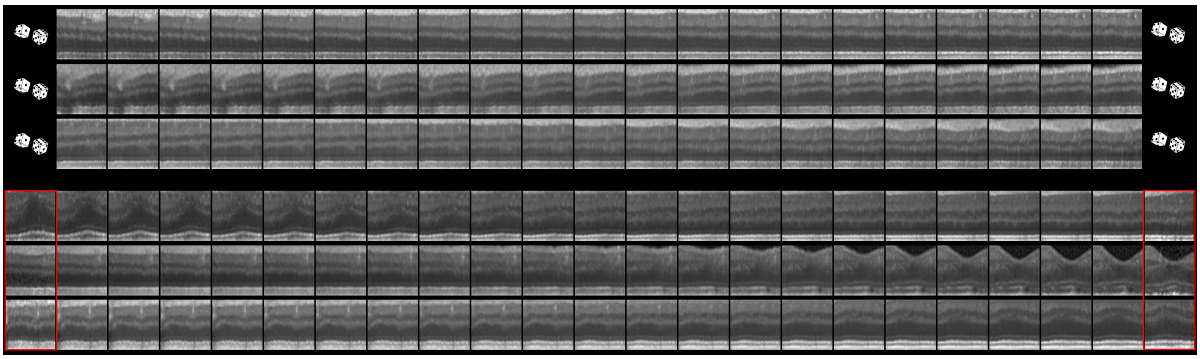

Figure 5: Interpolations in the z-space of the trained WGAN. First three rows: Linear interpolation in z-space between randomly sampled endpoints. Last three rows: Linear interpolation in z-space between two z locations conditioned by real images (images with red edgings) taken from the training set. (A more detailed visualization can be found in B.1.)

3.2. Results

The model captures normal variability in a smooth representation. In order to quantitatively evaluate the quality of generated images, we independently presented a visual Turing test to two retina experts. The accuracy of both raters for telling generated from real normal images apart was below random guessing (mean accuracy was 0.44), and the consensus between both raters was 0.58. These results clearly show that the model covers normal variability and images generated by our model are indistinguishable from real normal retinal OCT images even for clinical retina experts.

The mapping from real images to the latent space and back demonstrates that the model (1) learns a latent representation of normal anatomical variability, and the generator generates realistic looking images, (2) the proposed encoder training procedure enables to find the corresponding latent encoding for a given query image (see last three rows in Figure 5). The model is able to find a visually "close" normal image for a given input image not part of the training data (Figure 6). This is a prerequisite for anomaly detection. Both, sampling along lines between random positions in the z-space (see first three rows in Figure 5), or between positions resulting from encoding real images (see last three rows in Figure 5), show smooth transitions in z-space and the ability of our model to find very close "reconstructions" of real images. This suggests that WGAN training was successful, and that the WGAN captures the full range (visual variability) of normal images, presented during training.

The model accurately detects anomalous images. Table 1 presents an in-depth comparison on image-level anomaly detection accuracy of f-AnoGAN, joint encoder-decoder learning approaches (*AE*, *AdvAE*, and *ALI*), a direct utilization of the WGAN discriminator trained on normal images (A_D), and iterative z-mapping on a trained WGAN (*iterative*). Figure 8a shows corresponding ROC curves and AUC values. During inference, *f-AnoGAN* is as fast as the joint encoder-decoder learning approaches, and thus is much faster than the *iterative* approach (which most closely corresponds to *AnoGAN*). Most importantly, *f-AnoGAN* has higher accuracy than all other approaches. A direct utilization of the trained WGAN discriminator, A_D, achieves poor accuracy. Note that *iterative* results could be possibly improved towards reaching *f-AnoGAN* accuracy through more iterations during inference, but on a Titan-X GPU the current 300 updates for each of the 70,000 test images took in total 2 days to compute while *f-AnoGAN* took 20 seconds. Thus, the accuracy advantage of *f-AnoGAN* is also valid in practical terms.

In Figure B.3, the distributions of anomaly scores $\mathcal{A}(\mathbf{x})$ (according to Equation (4)) are independently plotted for normal and anomalous images of the test set. It underlines the discriminative ability of *f-AnoGAN* to differentiate between normal and anomalous images and supports the quantitative evaluation results of the *f-AnoGAN* model presented in Table 1.

The model localizes anomalies in images. We qualitatively compared the pixel-level anomaly localization performance of *f-AnoGAN* to *AE*, *AdvAE*, and *ALI*. The results are depicted in Figure 6. The first four columns show results on images used for WGAN training, the following four columns show results on normal images extracted from healthy cases of the test set, and the remainder columns show results on normal and anomalous images extracted from diseased cases of the test set.

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

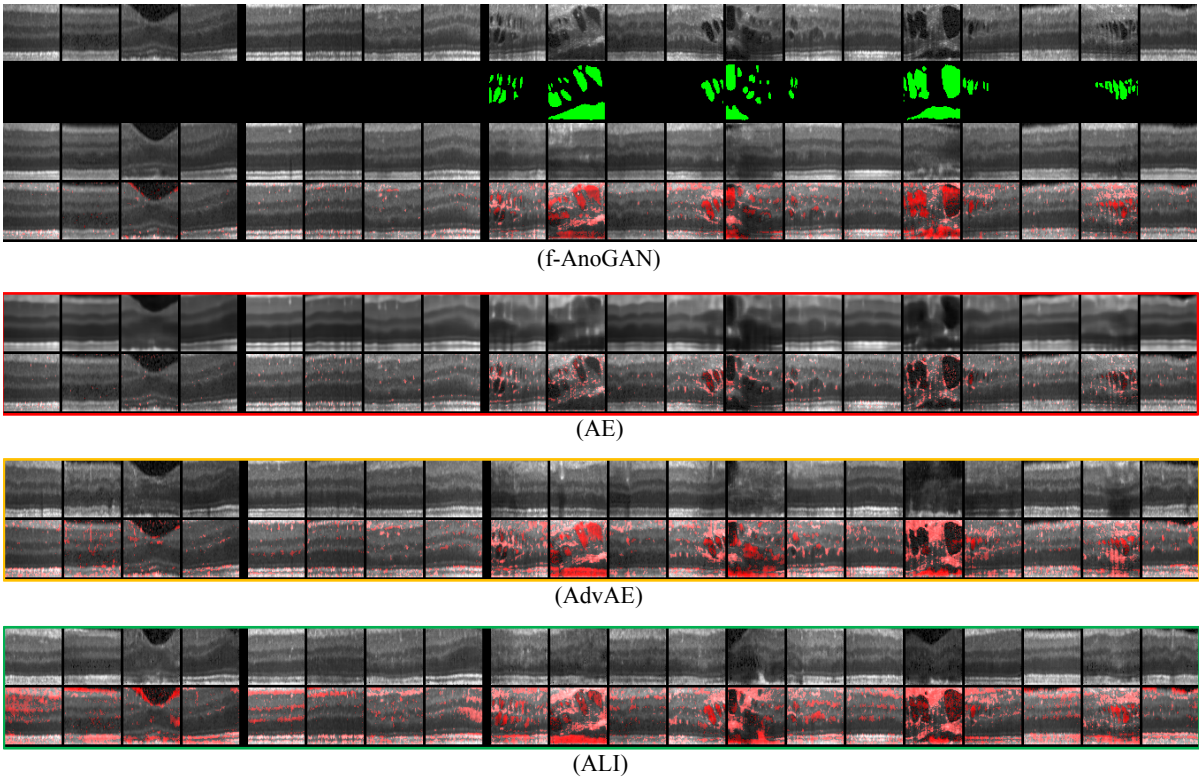

Figure 6: Model comparison on pixel-level localization of anomalous image regions. The first row shows real input images. Second row: Pixel-level anomaly (retinal fluid) annotations (only used for evaluation). First block: Images taken from the training set (normal images). Second block: Normal images extracted from healthy cases in the test set. Third block: Images extracted from diseased cases in the test set showing normal and anomalous patches. Starting with the third row, we show in one row the generated images and in subjacent rows the real query images with overlayed residual: Proposed *f-AnoGAN* (row 3 and 4), *AE* (row 5 and 6), *AdvAE* (row 7 and 8), and *ALI* (row 9 and 10). (Best viewed in color)

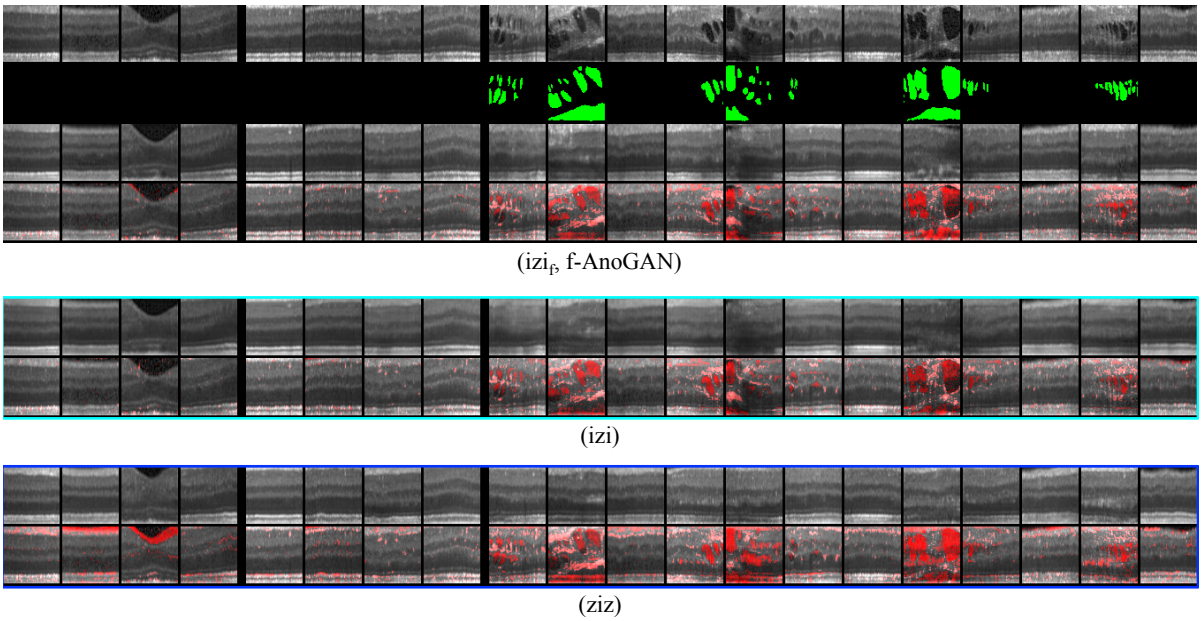

Figure 7: Encoder training comparison on pixel-level localization of anomalous image regions. Starting with the third row, we show in one row the generated images and in subjacent rows the real query images with overlayed residual: Proposed *f-AnoGAN* izi_f (row 3 and 4), *izi* (row 5 and 6), and *ziz* (row 7 and 8) encoder training. (Best viewed in color)

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

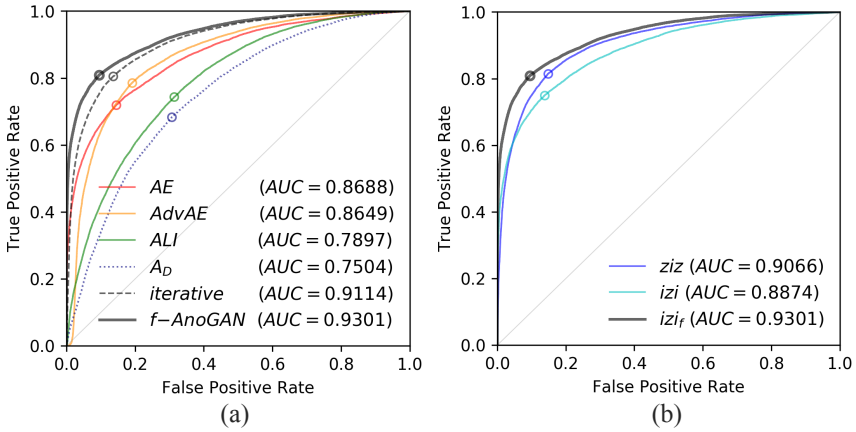

Figure 8: Image-level anomaly detection accuracy evaluation. Comparison of different approaches for image-level anomaly detection based on receiver operating characteristic (ROC) curves and corresponding area under the ROC curve (AUC) values (specified in parentheses). (a) Comparison of the proposed *f-AnoGAN* and alternative approaches: *AE* (red), *AdvAE* (yellow), *ALI* (green), directly utilizing the discriminator's output of the trained WGAN (A_D, blue dotted), *iterative* z-mapping based on a trained WGAN (following *AnoGAN* (Schlegl et al., 2017a), gray dashed), and the proposed *f-AnoGAN* model (gray). (b) Comparison of different encoder training approaches based on the same pre-trained WGAN: *ziz* encoder training (blue), *izi* encoder training (cyan), and izi_f encoder training (gray), which is implemented by the proposed *f-AnoGAN* model. (Please find more details on the different approaches in the main text.)

f-AnoGAN generates plausible normal versions of anomalous image regions, and correspondingly the residual is a good guide for segmentation. It is apparent when comparing with ground-truth annotations of known lesions (fluid-filled regions). Furthermore, *f-AnoGAN* does not show distinct false positive segmentation results on normal cases such as those occurring at the boundary between retina and vitreous with the *ALI* model or with the *ziz* encoder training approach (see third column of last row in Figure 6 and Figure 7). The *AE* model also does not show distinct false positive segmentations of normal cases, but is not able to segment anomalous image regions well. Both observations are explained by the fact that an *AE* of a comparable model architecture tends to overfit on local image content. Despite the fact that *AE* yields good image-level anomaly detection accuracy, this approach shows worst anomaly localization performance among all alternative approaches. In general, an *AE* generates over-smooth images for both inputs, normal and anomalous query images. In contrast, models implementing adversarial training generate more realistic images rich in detail. In comparison to remainder alternative approaches, the *AdvAE* shows good anomaly localization. However, in direct comparison to the *f-AnoGAN* model, the *AdvAE* shows slightly more false positives on normal images (most apparent in images of the training set), fails to localize relevant anomalies on the pixel level, and is more prone to oversegmenting regions. Despite the fact that the *ALI* model generates compelling realistic images, which have visual similarities to normal query images, these generated images and "reconstructions" for anomalous images show differences in the overall image content and thus in general yield to oversegmentations. We conclude, that the *f-AnoGAN* model not only yields best anomaly detection on image level but also qualitatively outperforms all investigated alternative approaches on pixel-level anomaly localization.

To provide a sense of the accuracy, for fluid "segmentation", the *f-AnoGAN* approach has a sensitivity of 0.69 and specificity of 0.75 with an AUC of 0.78. These sensitivity and specificity values are computed at the Youden index of the ROC curve, which is the optimal cut-off point that simultaneously maximizes sensitivity and specificity. A comparison on pixel-level anomaly localization accuracy of *f-AnoGAN*, *AE*, *AdvAE*, and *ALI* model reveals that *f-AnoGAN* outperforms the alternative approaches. Quantitative results and corresponding ROC curves are presented in Appendix B.2. Note that these numbers are only a coarse indication due to annotations cover only part of possible anomalies, and the corresponding unreliable amount of false positives.

Comparison of different encoder training architectures. Figure 8b shows ROC curves of the three investigated encoder training architectures. Quantitative results on the anomaly detection accuracy of the three encoder training approaches are summarized in Table 2. *ziz* encoder training yields slightly

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

Table 1: Clinical performance statistics calculated at the Youden index of the receiver operating characteristic (ROC) curve, the corresponding area under the ROC curve (AUC) and f-score measuring the image-level anomaly detection performance of a *convolutional autoencoder* (*AE*), *adversarial convolutional autoencoder* (*AdvAE*), *ALI* model, based on the output of the WGAN discriminator (A_D), iterative z-mapping utilizing the trained WGAN model (*iterative*) following Schlegl et al. (2017a), and our proposed *fast AnoGAN* (*f-AnoGAN*).

	Precision	Sensitivity	Specificity	f-score	AUC
AE	0.6824	0.7195	0.8550	0.7005	0.8688
AdvAE	0.6405	0.7856	0.8092	0.7057	0.8649
ALI	0.5063	0.7434	0.6863	0.6023	0.7897
A_D	0.4909	0.6831	0.6931	0.5713	0.7504
iterative	0.7202	0.8049	0.8645	0.7602	0.9114
f-AnoGAN	**0.7863**	**0.8091**	**0.9049**	**0.7975**	**0.9301**

Table 2: Comparison of investigated encoder training architectures: *ziz*, *izi* and izi_f (*f-AnoGAN*) based on the same WGAN training.

	Precision	Sensitivity	Specificity	f-score	AUC
ziz	0.7047	**0.8146**	0.8522	0.7557	0.9066
izi	0.7018	0.7497	0.8621	0.7250	0.8874
izi_f	**0.7863**	0.8091	**0.9049**	**0.7975**	**0.9301**

better results than *izi* encoder training. This can be explained by the fact that the *ziz* approach directly formulates the main problem of finding z locations. Adding a residual loss on the discriminator features of real and generated images in the izi_f encoder training architecture boosts the performance of the image-to-image mapping approach (also formulated in the *izi* training), so that this approach outperforms *ziz* encoder training. One explanation for the fact that the *ziz* approach does not yield the best results, lies obviously in the fact that *ziz* training does not "see" a single real image during training. Figure 7 depicts the ability of the three encoder training approaches to reconstruct query images and visually evaluates the anomaly localization ability. In comparison with izi_f, the *izi* encoder training architecture yields smooth "reconstructions", which has to be attributed to the objective function that solely minimizes mean image residuals but does not enforce realistic appearance of reconstructions. The izi_f and *ziz* encoder training approaches both yield realistic image reconstructions. Regarding the overall image content, the results of the *ziz* approach show a reduced similarity to query images, and thus are similar to "reconstructions" generated with the *ALI* approach. Both approaches are characterized by good coverage of anomalous regions by resulting segmentations but on the other hand *ziz* is more prone to oversegmentations.

So far, we examined the anomaly detection accuracy of the different encoder training approaches, when the same information is used for both steps encoder training and anomaly scoring. Results highlight that the utilization of the discriminator's features improves the final anomaly detection performance. The question remains, whether the utilization of the discriminator during anomaly scoring suffices or the utilization of the discriminator during encoder training helps to train a better encoder and consequently further improves anomaly detection results. To answer the question, we examined the anomaly detection accuracy of the *izi* and *ziz* encoder training approaches when during anomaly scoring the same – namely the full *f-AnoGAN* – anomaly score is used. Therefore, we utilized the discriminator (features) for anomaly scoring subsequent to *izi* and *ziz* encoder training. Results indicate, that the utilization of the discriminator during anomaly scoring (as used in the *f-AnoGAN* anomaly score) improves the anomaly detection accuracy regardless of the utilized preceding encoder training strategy. *izi* now yields a f-score and AUC of 0.7328 (0.7250) and 0.8976 (0.8874) respectively, and *ziz* yields a f-score and AUC of 0.7830 (0.7557) and 0.9209 (0.9066) respectively (results without the utilization of the discriminator during anomaly scoring are given in parenthesis). However, *f-AnoGAN*, which utilizes the discriminator already during encoder training, still outperforms the *izi* and *ziz* encoder training strategies.

4. Conclusion

We proposed fast anomaly detection using GANs and a related encoder training procedure. To create *f-AnoGAN*, we use healthy examples to train a WGAN and subsequently an encoder that maps images to the latent space for fast inference and anomaly detection. During anomaly detection, input images are reconstructed using the encoder and the generator, and a combined score of image reconstruction residual and a residual on discriminator features yields a reliable marker for anomalies. The resulting model outperforms alternative models on an anomaly detection task in retinal imaging data. It offers a significant acceleration during inference compared to the previously proposed AnoGAN, which relied on iterative estimation of latent space embeddings. The WGAN learns a smooth representation of the training data variability, and the encoder maps input images to positions that generate realistic (normal) replicas.

We investigated different encoder training approaches. A z-image-z encoder training architecture yields an encoder that maps images to latent encodings, which generate realistic reconstructed images when used in an image-z-image model during anomaly detection. Since the *ziz* encoder training objective operates in the latent space only, even normal query images are reconstructed with limited accuracy during anomaly detection. Consequently, it is less accurate when used for anomaly detection. An image-z-image model yields smooth reconstructions. Additionally including the WGAN discriminator in the encoder training yields the best anomaly detection and localization results. It creates reconstructions that are most similar to the query image.

We evaluated *f-AnoGAN* on OCT images of the retina. Unsupervised training was performed on OCT images of healthy subjects. In an experiment with two retina experts, we could show that the images generated by our model are indistinguishable from real normal retinal OCT images. It should be mentioned that in clinical practice OCT diagnosis is not performed at such small scale (64x64 pixels), but ophthalmologists are rather used to rate full-width OCT slices. Therefore, this task does not reflect the performance of the clinical retina experts on full OCT cross-sectional images. The automatic generation of full OCT slices is left for future research work. When applying the trained *f-AnoGAN* to new data, empirical results demonstrate good accuracy for detecting anomalous images. A limitation of the evaluation set-up is that only annotated lesions are used for the calculation of detection accuracy. Thus, "false positives" might be true anomalies not part of the annotated category.

The proposed methodology is generally applicable to anomaly detection on a variety of biomedical data such as 1D data (e.g. flow cytometry), time series, audio data, or 2D and 3D imaging data. It can be successfully applied whenever large amounts of normal medical (imaging) data are available for GAN training and the test distribution equals the training distribution. We find that the main area of application of *f-AnoGAN* lies in the distinction of normal and anomalous images (or volume scans) and providing corresponding hints in the form of coarse anomaly localizations, which define areas ("region proposals") that should be further assessed by clinical experts as candidates for disease markers or through a subsequent classification approach.

The speed of *f-AnoGAN* anomaly detection makes it a feasible approach for anomaly detection and large-throughput screening in practice.

When comparing AE and AdvAE, the AE over-adapts to anomalous image regions, which leads to worst anomaly localization performance. This can be attributed to the training objective of the AE only maximizing mean image similarity of input and reconstructed image. An AdvAE yields good anomaly localization performance but performs equally to the AE on anomaly detection. ALI shows overall good realistic image "reconstructions" but the query and "reconstructed" images also deviate in visual appearance, i.e. although they show similar semantic content deviate in their specific pixel-level appearance. The joint encoder-decoder learning approaches yield both, the image to latent space mapping and a generative model, in a single training step, which makes them - because of the simple training procedure - interesting approaches for real world applications. However, the *f-AnoGAN* model not only yields better pixel-level anomaly localization performance but also outperforms the alternative approaches on image-level anomaly detection.

Limitations. The quantitative evaluation of segmentation accuracy of anomaly detection only serves as a coarse indication to show that it can localize anomalies. Annotations cover only part of possible anomalies, and thus result in unreliable estimates of false positives and corresponding quality measures,

in particular when viewed in the light of biomarker discovery. Furthermore, the purpose of the method is primarily to detect anomalous images, and highlighted pixels should be viewed as coarse localizations of anomalies that could facilitate the identification of region proposals.

Acknowledgements

This work has received funding from the Austrian Federal Ministry of Science, Research and Economy (CDL OPTIMA), IBM (IBM Fellowship and Faculty Grant), Austrian Science Fund FWF (I2714-B31), and Austrian National Bank Anniversary Fund OeNB (15356, 15929). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for this research.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. URL http://tensorflow.org/

Akoglu, L., Tong, H., Koutra, D., 2015. Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery 29 (3), 626–688.

Alguliyev, R., Aliguliyev, R., Sukhostat, L., 2017. Anomaly detection in big data based on clustering. Statistics, Optimization & Information Computing 5 (4), 325–340.

Alipanahi, B., Delong, A., Weirauch, M. T., Frey, B. J., 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology 33 (8), 831.

Anbeek, P., Vincken, K. L., van Osch, M. J., Bisschops, R. H., van der Grond, J., 2004. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. Medical image analysis 8 (3), 205–215.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv preprint arXiv:1701.07875.

Awoyemi, J. O., Adetunmbi, A. O., Oluwadare, S. A., 2017. Credit card fraud detection using machine learning techniques: A comparative analysis. In: International Conference on Computing Networking and Informatics (ICCNI), 2017. IEEE, pp. 1–9.

Barratt, S., Sharma, R., 2018. A Note on the Inception Score. arXiv preprint arXiv:1801.01973.

Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B., Langlotz, C. P., Amrhein, T. J., Lungren, M. P., 2017. Deep learning to classify radiology free-text reports. Radiology, 171115.

Chuquicusma, M. J., Hussein, S., Burt, J., Bagci, U., 2018. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on. IEEE, pp. 240–244.

Cireşan, D. C., Giusti, A., Gambardella, L. M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 411–418.

Creswell, A., Bharath, A. A., 2018. Inverting The Generator Of A Generative Adversarial Network (II). NIPS 2016 Workshop on Adversarial Training. arXiv preprint arXiv:1802.05701.

Donahue, J., Krähenbühl, P., Darrell, T., 2016. Adversarial feature learning. arXiv:1605.09782.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., Courville, A., 2016. Adversarially learned inference. arXiv preprint arXiv:1606.00704.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., Leckie, C., 2016. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition 58, 121–134.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115.

Farsiu, S., Chiu, S. J., Izatt, J. A., Toth, C. A., 2008. Fast detection and segmentation of drusen in retinal optical coherence tomography images. In: Ophthalmic Technologies XVIII. Vol. 6844. International Society for Optics and Photonics, p. 68440D.

Garvin, M. K., Abràmoff, M. D., Wu, X., Russell, S. R., Burns, T. L., Sonka, M., 2009. Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. Transactions on Medical Imaging, IEEE 28 (9), 1436–1447.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680.

Grewal, M., Srivastava, M. M., Kumar, P., Varadarajan, S., 2017. RADNET: Radiologist Level Accuracy using Deep Learning for HEMORRHAGE detection in CT Scans. arXiv preprint arXiv:1710.04934.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein GANs. arXiv preprint arXiv:1704.00028.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.

Hinton, G., Srivastava, N., Swersky, K., 2013. *Rmsprop: Divide the gradient by a running average of its recent magnitude.* Coursera: Neural Networks for Machine Learning. Lecture 6e. Accessed September 23^{th} 2018, `https://www.coursera.org/lecture/neural-networks/rmsprop-divide-the-gradient-by-a-running-average-of-its-recent-magnitude-YQHki`.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 597–609.

Kaus, M. R., Warfield, S. K., Nabavi, A., Black, P. M., Jolesz, F. A., Kikinis, R., 2001. Automated segmentation of MR images of brain tumors. Radiology 218 (2), 586–591.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312.

Lipton, Z. C., Tripathi, S., 2017. Precise recovery of latent vectors from generative adversarial networks. ICLR 2017 workshop track. arXiv preprint arXiv:1702.04782.

Miotto, R., Li, L., Kidd, B. A., Dudley, J. T., 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports 6, 26094.

Nguyen, L. H., Goulet, J.-A., 2017. Anomaly detection with the switching kalman filter for structural health monitoring. Structural Control and Health Monitoring.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A. A., 2016. Context encoders: Feature learning by inpainting. CoRR abs/1604.07379.
URL `http://arxiv.org/abs/1604.07379`

Pimentel, M. A., Clifton, D. A., Clifton, L., Tarassenko, L., 2014. A review of novelty detection. Signal Processing 99, 215–249.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225.

Ramteke, R., Monali, Y. K., 2012. Automatic medical image classification and abnormality detection using k-nearest neighbour. International Journal of Advanced Computer Research 2 (4), 190–196.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N., 2017. Abnormal event detection in videos using generative adversarial nets. In: IEEE International Conference on Image Processing (ICIP).

Reddick, W. E., Glass, J. O., Cook, E. N., Elkin, T. D., Deaton, R. J., 1997. Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks. IEEE Transactions on medical imaging 16 (6), 911–918.

Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E., 2018. Adversarially learned one-class classifier for novelty detection. arXiv preprint arXiv:1802.09088.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. In: Advances in Neural Information Processing Systems. pp. 2226–2234.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., Langs, G., 2017a. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 146–157.

Schlegl, T., Waldstein, S. M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A.-M., Podkowinski, D., Gerendas, B. S., Langs, G., Schmidt-Erfurth, U., 2017b. Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology.

Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., Langs, G., 2015. Predicting semantic descriptions from medical images with convolutional neural networks. In: International Conference on Information Processing in Medical Imaging. Vol. 24. Springer, pp. 437–448.

Schmidt-Erfurth, U., Bogunovic, H., Sadeghipour, A., Schlegl, T., Langs, G., Gerendas, B. S., Osborne, A., Waldstein, S. M., 2018. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. Ophthalmology Retina 2 (1), 24–30.

Seeböck, P., Waldstein, S. M., Klimscha, S., Bogunovic, H., Schlegl, T., Gerendas, B. S., Donner, R., Schmidt-Erfurth, U., Langs, G., 2018. Unsupervised Identification of Disease Marker Candidates in Retinal OCT Imaging Data. IEEE Transactions on medical imaging.

Sidibe, D., Sankar, S., Lemaitre, G., Rastgoo, M., Massich, J., Cheung, C. Y., Tan, G. S., Milea, D., Lamoureux, E., Wong, T. Y., et al., 2017. An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images. Computer methods and programs in biomedicine 139, 109–117.

Stansfield, S. A., 1986. Angy: A rule-based expert system for automatic segmentation of coronary vessels from digital subtracted angiograms. IEEE Transactions on Pattern Analysis and Machine Intelligence (2), 188–199.

Tandiya, N., Jauhar, A., Marojevic, V., Reed, J. H., 2018. Deep Predictive Coding Neural Network for RF Anomaly Detection in Wireless Networks. arXiv preprint arXiv:1803.06054.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE transactions on medical imaging 18 (10), 897–908.

Venhuizen, F. G., van Ginneken, B., Bloemen, B., van Grinsven, M. J., Philipsen, R., Hoyng, C., Theelen, T., Sánchez, C. I., 2015. Automated age-related macular degeneration classification in OCT using unsupervised feature learning. In: SPIE Medical Imaging. Vol. 9414. International Society for Optics and Photonics, p. 94141I.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proc. of the 25th ICML. ACM, pp. 1096–1103.

Walsh, S. L., Calandriello, L., Sverzellati, N., Wells, A. U., Hansell, D. M., 2015. Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. Thorax, thoraxjnl–2015.

Wang, Z., Hauser, N., Singer, G., Trippel, M., Kubik-Huch, R. A., Schneider, C. W., Stampanoni, M., 2014. Non-invasive classification of microcalcifications with phase-contrast X-ray mammography. Nature communications 5, 3797.

Weisenfeld, N. I., Mewes, A. U., Warfield, S. K., 2006. Highly accurate segmentation of brain tissue and subcortical gray matter from newborn MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 199–206.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., Chandrasekhar, V. R., 2018. Efficient GAN-Based Anomaly Detection. arXiv preprint arXiv:1802.06222.

Zhang, L., Lin, J., Karim, R., 2018. Adaptive kernel density-based anomaly detection for nonlinear systems. Knowledge-Based Systems 139, 50–63.

Zheng, Y.-J., Zhou, X.-H., Sheng, W.-G., Xue, Y., Chen, S.-Y., 2018. Generative adversarial network based telecom fraud detection at the receiving bank. Neural Networks.

Zwirewich, C. V., Vedal, S., Miller, R. R., Müller, N. L., 1991. Solitary pulmonary nodule: high-resolution CT and radiologic-pathologic correlation. Radiology 179 (2), 469–476.

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

Appendix A. Additional experiments

Appendix A.1. Denoising encoder training

The encoder training approaches (Section 2.2) share architectural similarities with autoencoder (AE) training and the encoder training network has more model parameters than inputs. Hence, the network is prone to learn the "identity function" instead of learning to differentiate samples drawn from the training distribution and any other input configuration. Therefore, we additionally examined whether applying input noise, following the *denoising AE* approach as suggested by Vincent et al. (2008), beneficially influences learning the mapping from images to latent vector encodings.

Vincent et al. (2008) suggested the following input corruption process: a fixed amount of pixel values at randomly chosen positions are set to 0. Vincent et al. (2008) experimented with different percentages of input pixels up to 50%. We set 20% of input pixels at randomly chosen positions to 0, but found already at this amount that applying noise on the inputs during encoder training tends to yield smooth reconstructions, which would especially be disadvantageous on normal imaging data comprising high textural diversity as for instance on chest computed tomography (CT) data. Smooth reconstructions lead in further consequence to erroneous anomaly detection in normal images (i.e. false positives).

Figure A.1 shows exemplarily pixel-based anomaly localization results on normal images of the test set. We suspect that a denoising approach does not work that well for the encoder training approach, since - unlike in the standard AE training - we only learn the encoder but keep the decoder fixed. Therefore we did not apply input noise in the presented encoder training approaches and leave the assessment of the general influence of input noise on various imaging data with various corruption procedures to future studies.

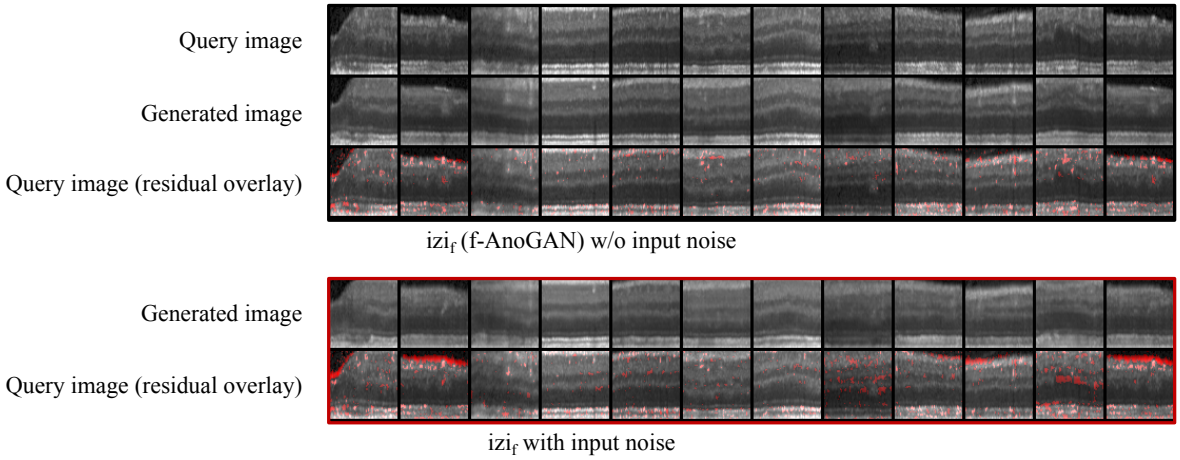

Figure A.1: Qualitative encoder training comparison on normal test images not used during training. In general, the (izi_f) encoder trained without applying noise on the inputs reconstructs query images more accurately. Whereas denoising encoder training (i.e. applying noise on the inputs during training) tends to generate smooth reconstructions of the query images that also show more marked deviations regarding global appearance and thus yield to higher false positive rates. (Best viewed in color)

Appendix A.2. Constraining encoder outputs

During WGAN training z locations are sampled from a normal distribution, and thus a large proportion of latent encodings of normal images is placed in its central part. The generator has more opportunity to improve on generating images from those locations able to fool the discriminator. Hence these are images, for which the generator receives most indirect feedback from the discriminator, namely images that share main components of visual variability. During encoder training, we restrict the encoder to map normal images into the range $(-1\sigma, +1\sigma)$ of the standard normal distribution by implementing a *tanh* activation function on the output layer of the encoder. This form of constraining the encoder is only possible when using a normal distribution (as opposed to a uniform distribution) for sampling z locations during WGAN training. Applying *tanh* activation function on the outputs of the last encoder

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

layer, which represent the latent vectors, consistently yields a gain in accuracy for all of the examined encoder architectures in direct comparison to an utilization of a linear output layer. Qualitative results are depicted in Figure A.2 and quantitative results for the ziz, izi, and izi_f encoder architectures are presented in Table A.1. The ziz architecture benefits most, whereas in the izi_f encoder architecture, which is utilized in our proposed *f-AnoGAN* model, we observe only a small gain in accuracy. Since the unconstrained izi_f encoder training approach already outperforms the constrained ziz and izi encoder training approaches, this result substantiates the superiority of the izi_f encoder training approach. Since even for the best performing izi_f encoder architecture, the utilization of a *tanh* activation function improves the performance to some extent, we suggest to use this constraint on encoder outputs. Note that these are preliminary empirical results without theoretical proof.

Table A.1: Comparison of studied encoder training approaches: ziz, izi and izi_f implementing an unconstrained encoder through a linear output layer based on the same WGAN training. Performance outcomes for the *tanh*-version are given in parenthesis.

	Precision	Sensitivity	Specificity	f-score	AUC
ziz	0.6412 (0.7047)	0.8018 (0.8146)	0.8058 (0.8522)	0.7126 (0.7557)	0.8848 (0.9066)
izi	0.6955 (0.7018)	0.7265 (0.7497)	0.8623 (0.8621)	0.7107 (0.7250)	0.8752 (0.8874)
izi_f	**0.7692** (0.7863)	**0.8110** (0.8091)	**0.8947** (0.9049)	**0.7896** (0.7975)	**0.9269** (0.9301)

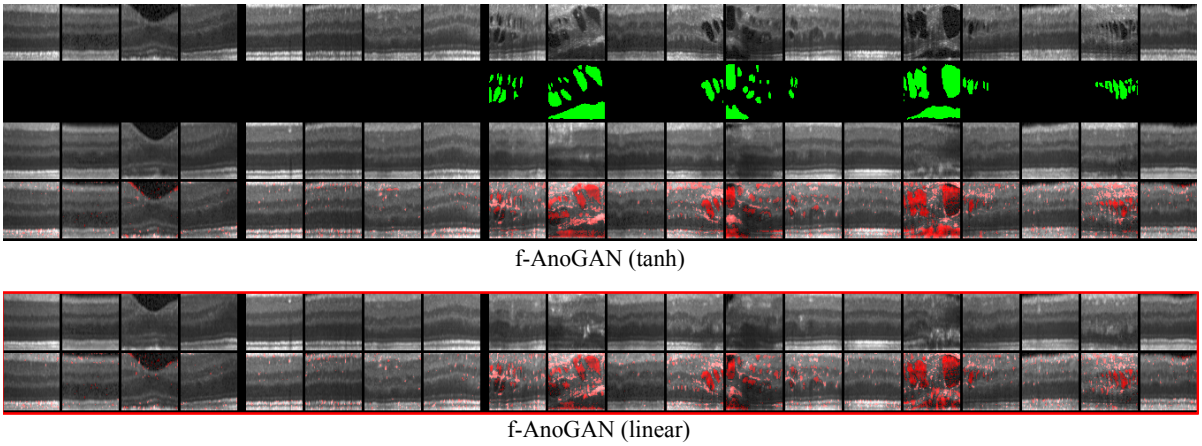

Figure A.2: Qualitative anomaly localization performance comparison. Pixel-level anomaly detection resulting from encoder training using a *tanh* activation function on the encoder output versus an unconstrained encoder implementing a linear output layer. (Best viewed in color)

Appendix B. Additional results

Appendix B.1. Interpolations in the z-space

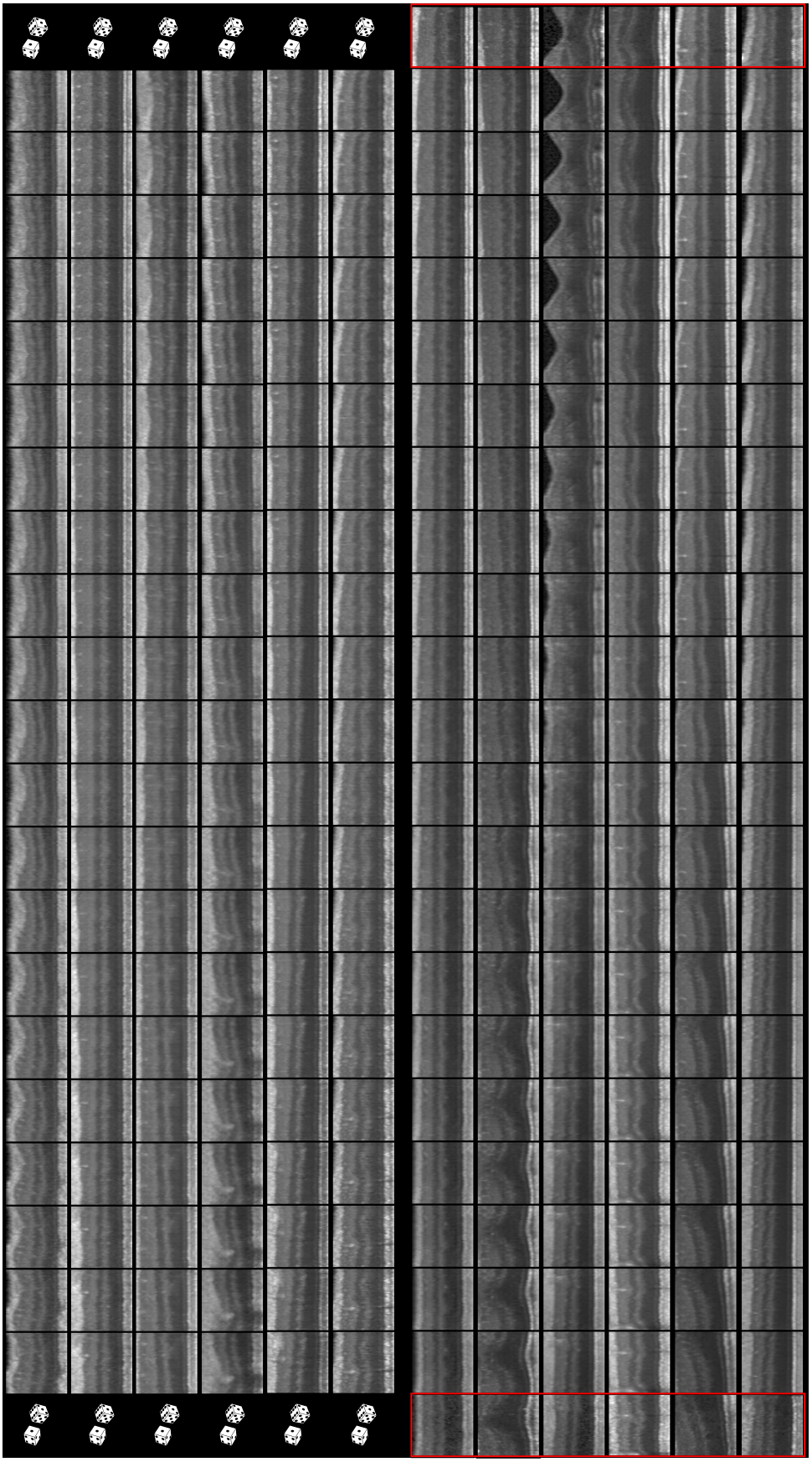

Figure B.1: Interpolations in the z-space of the trained WGAN. First six rows: Linear interpolation in z-space between randomly sampled endpoints. Last six rows: Linear interpolation in z-space between two z locations conditioned by real images (images with red edgings) taken from the training set.

Published in Medical Image Analysis (2019). DOI: https://doi.org/10.1016/j.media.2019.01.010.

Appendix B.2. Pixel-level anomaly localization

Table B.1: Clinical performance statistics calculated at the Youden index of the receiver operating characteristic (ROC) curve, and the corresponding area under the ROC curve (AUC) measuring the pixel-level anomaly localization performance of a *convolutional autoencoder* (*AE*), *adversarial convolutional autoencoder* (*AdvAE*), *ALI* model, and our proposed *fast AnoGAN* (*f-AnoGAN*).

	Sensitivity	Specificity	AUC
AE	**0.6960**	0.5428	0.6459
AdvAE	0.6284	0.7157	0.7195
ALI	0.6043	0.6468	0.6610
f-AnoGAN	0.6907	**0.7534**	**0.7831**

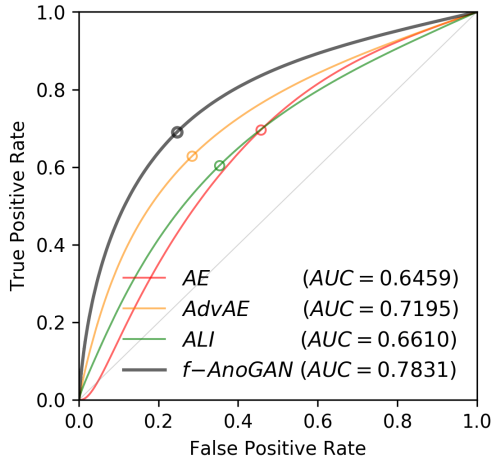

Figure B.2: Pixel-level anomaly localization accuracy evaluation. Comparison of different approaches based on receiver operating characteristic (ROC) curves and corresponding area under the ROC curve (AUC) values (specified in parentheses): *AE* (red), *AdvAE* (yellow), *ALI* (green), and the proposed *f-AnoGAN* model (gray). (Best viewed in color)

Appendix B.3. Distribution of anomaly scores

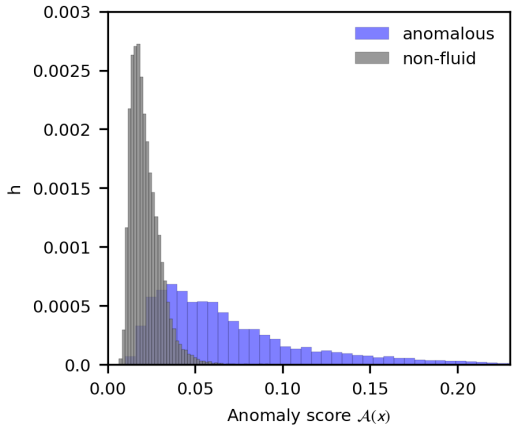

Figure B.3: Discrete distributions of *f-AnoGAN* anomaly scores computed for the full test set that comprises 70,000 images. Relative frequency histograms of anomaly scores are independently plotted for images that contain retinal fluid ("anomalous") and for images that do not contain retinal fluid ("non-fluid"). The images showing retinal fluid are obviously expected to be scored as *anomalous* images. Since for training only non-fluid images were used, the non-fluid images of the test set most likely have to be rated as *normal* images. However, these images may also contain anomalies, which were not present in the training set. (Best viewed in color)