

# Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery

Thomas Schlegl<sup>1,2</sup> \*, Philipp Seeböck<sup>1,2</sup>, Sebastian M. Waldstein<sup>2</sup>,  
Ursula Schmidt-Erfurth<sup>2</sup>, and Georg Langs<sup>1</sup>

<sup>1</sup>Computational Imaging Research Lab, Department of Biomedical Imaging and  
Image-guided Therapy, Medical University Vienna, Austria

thomas.schlegl@meduniwien.ac.at

<sup>2</sup>Christian Doppler Laboratory for Ophthalmic Image Analysis, Department of  
Ophthalmology and Optometry, Medical University Vienna, Austria

**Abstract.** Obtaining models that capture imaging markers relevant for disease progression and treatment monitoring is challenging. Models are typically based on large amounts of data with annotated examples of known markers aiming at automating detection. High annotation effort and the limitation to a vocabulary of known markers limit the power of such approaches. Here, we perform unsupervised learning to identify anomalies in imaging data as candidates for markers. We propose *AnoGAN*, a deep convolutional generative adversarial network to learn a manifold of normal anatomical variability, accompanying a novel anomaly scoring scheme based on the mapping from image space to a latent space. Applied to new data, the model labels anomalies, and scores image patches indicating their fit into the learned distribution. Results on optical coherence tomography images of the retina demonstrate that the approach correctly identifies anomalous images, such as images containing retinal fluid or hyperreflective foci.

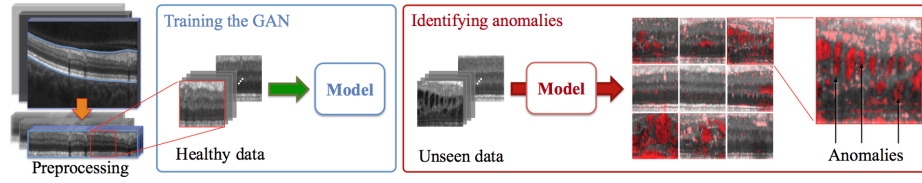
## 1 Introduction

The detection and quantification of disease markers in imaging data is critical during diagnosis, and monitoring of disease progression, or treatment response. Relying on the vocabulary of known markers limits the use of imaging data containing far richer relevant information. Here, we demonstrate that relevant *anomalies* can be identified by unsupervised learning on large-scale imaging data.

Medical imaging enables the observation of markers correlating with disease status, and treatment response. While there is a wide range of known markers (e.g., characteristic image appearance of brain tumors or calcification patterns in breast screening), many diseases lack a sufficiently broad set, while in others the predictive power of markers is limited. Furthermore, even if predictive

---

\* This work has received funding from IBM, FWF (I2714-B31), OeNB (15356, 15929), the Austrian Federal Ministry of Science, Research and Economy (CDL OPTIMA).



**Fig. 1.** Anomaly detection framework. The preprocessing step includes extraction and flattening of the retinal area, patch extraction and intensity normalization. Generative adversarial training is performed on healthy data and testing is performed on both, unseen healthy cases and anomalous data.

markers are known, their computational detection in imaging data typically requires extensive supervised training using large amounts of annotated data such as labeled lesions. This limits our ability to exploit imaging data for treatment decisions.

Here, we propose unsupervised learning to create a rich generative model of healthy local anatomical appearance. We show how generative adversarial networks (GANs) can solve the central problem of creating a sufficiently representative model of appearance, while at the same time learning a generative and discriminative component. We propose an improved technique for mapping from image space to latent space. We use both components to differentiate between observations that conform to the training data and such data that does not fit.

*Related Work* Anomaly detection is the task of identifying test data not fitting the *normal* data distribution seen during training. Approaches for anomaly detection exist in various domains, ranging from video analysis [1] to remote sensing [2]. They typically either use an explicit representation of the distribution of normal data in a feature space, and determine outliers based on the local density at the observations' position in the feature space. Carrera et al. [3] utilized convolutional sparse models to learn a dictionary of filters to detect anomalous regions in texture images. Erfani et al. [4] proposed a hybrid model for unsupervised anomaly detection that uses a one-class support vector machine (SVM). The SVM was trained from features that were learned by a deep belief network (DBN). The experiments in the aforementioned works were performed on real-life-datasets comprising 1D inputs, synthetic data or texture images, which have lower dimensionality or different data characteristics compared to medical images. An investigation of anomaly detection research papers can be found in [5]. In clinical optical coherence tomography (OCT) scan analysis, Venhuizen et al. [6] used bag-of-word features as a basis for supervised random forest classifier training to distinguish diseased patients from healthy subjects. Schlegl et al. [7] utilized convolutional neural networks to segment retinal fluid regions in OCT data via weakly supervised learning based on semantic descriptions of pathology-location pairs extracted from medical reports. In contrast to our approach, both works used some form of supervision for classifier training. Seeböck et al. [8] identified anomalous regions in OCT images through unsupervised learning on healthy examples, using a convolutional autoencoder and a

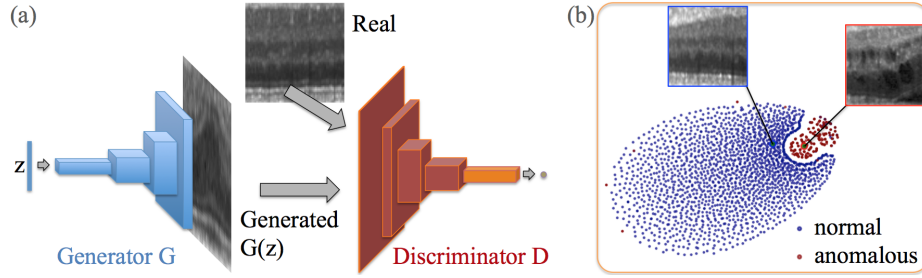
one-class SVM, and explored different classes of anomalies. In contrast to this work, the SVM in [8] involved the need to choose a hyper-parameter that defined the amount of training points covered by the estimated healthy region.

GANs enable to learn generative models generating detailed realistic images [9,10,11]. Radford et al. [12] introduced deep convolutional generative adversarial networks (DCGANs) and showed that GANs are capable of capturing semantic image content enabling vector arithmetic for visual concepts. Yeh et al. [13] trained GANs on natural images and applied the trained model for semantic image inpainting. Compared to Yeh et al. [13], we implement two adaptations for an improved mapping from images to the latent space. We condition the search in the latent space on the whole query image, and propose a novel variant to guide the search in the latent space (inspired by feature matching [14]). In addition, we define an anomaly score, which is not needed in an inpainting task. The main difference of this paper to aforementioned anomaly detection work is the representative power of the generative model and the coupled mapping schema, which utilizes a trained DCGAN and enables accurate discrimination between normal anatomy, and local anomalous appearance. This renders the detection of subtle anomalies at scale feasible.

*Contribution* In this paper, we propose adversarial training of a generative model of normal appearance (see blue block in Figure 1), described in Section 2.1, and a coupled mapping schema, described in Section 2.2, that enables the evaluation of novel data (Section 2.3) to identify anomalous images and segment anomalous regions within imaging data (see red block in Figure 1). Experiments on labeled test data, extracted from spectral-domain OCT (SD-OCT) scans, show that this approach identifies known anomalies with high accuracy, and at the same time detects other anomalies for which no voxel-level annotations are available. To the best of our knowledge, this is the first work, where GANs are used for anomaly or novelty detection. Additionally, we propose a novel mapping approach, wherewith the pre-image problem can be tackled.

## 2 Generative Adversarial Representation Learning to Identify Anomalies

To identify anomalies, we learn a model representing normal anatomical variability based on GANs [13]. This method trains a generative model, and a discriminator to distinguish between generated and real data simultaneously (see Figure 2(a)). Instead of a single cost function optimization, it aims at the Nash equilibrium of costs, increasing the representative power and specificity of the generative model, while at the same time becoming more accurate in classifying real- from generated data and improving the corresponding feature mapping. In the following we explain how to build this model (Section 2.1), and how to use it to identify appearance not present in the training data (Sections 2.2 and 2.3).



**Fig. 2.** (a) Deep convolutional generative adversarial network. (b) t-SNE embedding of normal (blue) and anomalous (red) images on the feature representation of the last convolution layer (orange in (a)) of the discriminator.

## 2.1 Unsupervised Manifold Learning of Normal Anatomical Variability

We are given a set of  $M$  medical images  $\mathbf{I}_m$  showing healthy anatomy, with  $m = 1, 2, \dots, M$ , where  $\mathbf{I}_m \in \mathbb{R}^{a \times b}$  is an intensity image of size  $a \times b$ . From each image  $\mathbf{I}_m$ , we extract  $K$  2D image patches  $x_{k,m}$  of size  $c \times c$  from randomly sampled positions resulting in data  $\mathbf{x} = x_{k,m} \in \mathcal{X}$ , with  $k = 1, 2, \dots, K$ . During training we are only given  $\langle \mathbf{I}_m \rangle$  and train a generative adversarial model to learn the manifold  $\mathcal{X}$  (blue region in Figure 2(b)), which represents the variability of the training images, in an unsupervised fashion. For testing, we are given  $\langle \mathbf{y}_n, l_n \rangle$ , where  $\mathbf{y}_n$  are unseen images of size  $c \times c$  extracted from new testing data  $\mathbf{J}$  and  $l_n \in \{0, 1\}$  is an array of binary image-wise ground-truth labels, with  $n = 1, 2, \dots, N$ . These labels are only given during testing, to evaluate the anomaly detection performance based on a given pathology.

*Encoding Anatomical Variability with a Generative Adversarial Network.* A GAN consists of two adversarial modules, a generator  $G$  and a discriminator  $D$ . The generator  $G$  learns a distribution  $p_g$  over data  $\mathbf{x}$  via a mapping  $G(\mathbf{z})$  of samples  $\mathbf{z}$ , 1D vectors of uniformly distributed input noise sampled from latent space  $\mathcal{Z}$ , to 2D images in the image space manifold  $\mathcal{X}$ , which is populated by healthy examples. In this setting, the network architecture of the generator  $G$  is equivalent to a convolutional decoder that utilizes a stack of strided convolutions. The discriminator  $D$  is a standard CNN that maps a 2D image to a single scalar value  $D(\cdot)$ . The discriminator output  $D(\cdot)$  can be interpreted as probability that the given input to the discriminator  $D$  was a real image  $\mathbf{x}$  sampled from training data  $\mathcal{X}$  or generated  $G(\mathbf{z})$  by the generator  $G$ .  $D$  and  $G$  are simultaneously optimized through the following two-player minimax game with value function  $V(G, D)$  [9]:

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

The discriminator is trained to maximize the probability of assigning real training examples the “real” and samples from  $p_g$  the “fake” label. The generator  $G$

is simultaneously trained to fool  $D$  via minimizing  $V(G) = \log(1 - D(G(\mathbf{z})))$ , which is equivalent to maximizing

$$V(G) = D(G(\mathbf{z})). \quad (2)$$

During adversarial training the generator improves in generating realistic images and the discriminator progresses in correctly identifying real and generated images.

## 2.2 Mapping new Images to the Latent Space

When adversarial training is completed, the generator has learned the mapping  $G(\mathbf{z}) = \mathbf{z} \mapsto \mathbf{x}$  from latent space representations  $\mathbf{z}$  to realistic (normal) images  $\mathbf{x}$ . But GANs do not automatically yield the inverse mapping  $\mu(\mathbf{x}) = \mathbf{x} \mapsto \mathbf{z}$  for free. The latent space has smooth transitions [12], so sampling from two points close in the latent space generates two visually similar images. Given a query image  $\mathbf{x}$ , we aim to find a point  $\mathbf{z}$  in the latent space that corresponds to an image  $G(\mathbf{z})$  that is visually most similar to query image  $\mathbf{x}$  and that is located on the manifold  $\mathcal{X}$ . The degree of similarity of  $\mathbf{x}$  and  $G(\mathbf{z})$  depends on to which extent the query image follows the data distribution  $p_g$  that was used for training of the generator. To find the best  $\mathbf{z}$ , we start with randomly sampling  $\mathbf{z}_1$  from the latent space distribution  $\mathcal{Z}$  and feed it into the trained generator to get a generated image  $G(\mathbf{z}_1)$ . Based on the generated image  $G(\mathbf{z}_1)$  we define a loss function, which provides gradients for the update of the coefficients of  $\mathbf{z}_1$  resulting in an updated position in the latent space,  $\mathbf{z}_2$ . In order to find the most similar image  $G(\mathbf{z}_\Gamma)$ , the location of  $\mathbf{z}$  in the latent space  $\mathcal{Z}$  is optimized in an iterative process via  $\gamma = 1, 2, \dots, \Gamma$  backpropagation steps.

In the spirit of [13], we define a loss function for the mapping of new images to the latent space that comprises two components, a *residual loss* and a *discrimination loss*. The *residual loss* enforces the visual similarity between the generated image  $G(\mathbf{z}_\gamma)$  and query image  $\mathbf{x}$ . The *discrimination loss* enforces the generated image  $G(\mathbf{z}_\gamma)$  to lie on the learned manifold  $\mathcal{X}$ . Therefore, both components of the trained GAN, the discriminator  $D$  and the generator  $G$ , are utilized to adapt the coefficients of  $\mathbf{z}$  via backpropagation. In the following, we give a detailed description of both components of the loss function.

**Residual Loss** The *residual loss* measures the visual dissimilarity between query image  $\mathbf{x}$  and generated image  $G(\mathbf{z}_\gamma)$  in the image space and is defined by

$$\mathcal{L}_R(\mathbf{z}_\gamma) = \sum |\mathbf{x} - G(\mathbf{z}_\gamma)|. \quad (3)$$

Under the assumption of a perfect generator  $G$  and a perfect mapping to latent space, for an ideal normal query case, images  $\mathbf{x}$  and  $G(\mathbf{z}_\gamma)$  are identical. In this case, the *residual loss* is zero.

**Discrimination Loss** For image inpainting, Yeh et al. [13] based the computation of the *discrimination loss*  $\mathcal{L}_{\hat{D}}(\mathbf{z}_\gamma)$  on the discriminator output by feeding the generated image  $G(\mathbf{z}_\gamma)$  into the discriminator  $\mathcal{L}_{\hat{D}}(\mathbf{z}_\gamma) = \sigma(D(G(\mathbf{z}_\gamma)), \alpha)$ , where  $\sigma$  is the sigmoid cross entropy, which defined the discriminator loss of real images during adversarial training, with logits  $D(G(\mathbf{z}_\gamma))$  and targets  $\alpha = 1$ .

**An improved discrimination loss based on feature matching** In contrast to the work of Yeh et al. [13], where  $\mathbf{z}_\gamma$  is updated to fool  $D$ , we define an alternative discrimination loss  $\mathcal{L}_D(\mathbf{z}_\gamma)$ , where  $\mathbf{z}_\gamma$  is updated to match  $G(\mathbf{z}_\gamma)$  with the learned distribution of normal images. This is inspired by the recently proposed feature matching technique [14].

Feature matching addresses the instability of GANs due to over-training on the discriminator response [14]. In the feature matching technique, the objective function for optimizing the generator is adapted to improve GAN training. Instead of optimizing the parameters of the generator via maximizing the discriminator’s output on generated examples (Eq. (2)), the generator is forced to generate data that has similar statistics as the training data, i.e. whose intermediate feature representation is similar to those of real images. Salimans et al. [14] found that feature matching is especially helpful when classification is the target task. Since we do not use any labeled data during adversarial training, we do not aim for learning class-specific discriminative features but we aim for learning good representations. Thus, we do not adapt the training objective of the generator during adversarial training, but instead use the idea of feature matching to improve the mapping to the latent space. Instead of using the scalar output of the discriminator for computing the *discrimination loss*, we propose to use a richer intermediate feature representation of the discriminator and define the *discrimination loss* as follows:

$$\mathcal{L}_D(\mathbf{z}_\gamma) = \sum |\mathbf{f}(\mathbf{x}) - \mathbf{f}(G(\mathbf{z}_\gamma))|, \quad (4)$$

where the output of an intermediate layer  $f(\cdot)$  of the discriminator is used to specify the statistics of an input image. Based on this new loss term, the adaptation of the coordinates of  $\mathbf{z}$  does not only rely on a hard decision of the trained discriminator, whether or not a generated image  $G(\mathbf{z}_\gamma)$  fits the learned distribution of normal images, but instead takes the rich information of the feature representation, which is learned by the discriminator during adversarial training, into account. In this sense, our approach utilizes the trained discriminator not as classifier but as a feature extractor.

For the mapping to the latent space, we define the overall loss as weighted sum of both components:

$$\mathcal{L}(\mathbf{z}_\gamma) = (1 - \lambda) \cdot \mathcal{L}_R(\mathbf{z}_\gamma) + \lambda \cdot \mathcal{L}_D(\mathbf{z}_\gamma). \quad (5)$$

Only the coefficients of  $\mathbf{z}$  are adapted via backpropagation. The trained parameters of the generator and discriminator are kept fixed.

### 2.3 Detection of Anomalies

During anomaly identification in new data we evaluate the new query image  $\mathbf{x}$  as being a normal or anomalous image. Our loss function (Eq. (5)), used for mapping to the latent space, evaluates in every update iteration  $\gamma$  the compatibility of generated images  $G(\mathbf{z}_\gamma)$  with images, seen during adversarial training. Thus, an *anomaly score*, which expresses the fit of a query image  $\mathbf{x}$  to the model of normal images, can be directly derived from the mapping loss function (Eq. (5)):

$$A(\mathbf{x}) = (1 - \lambda) \cdot R(\mathbf{x}) + \lambda \cdot D(\mathbf{x}), \quad (6)$$

where the *residual score*  $R(x)$  and the *discrimination score*  $D(x)$  are defined by the *residual loss*  $\mathcal{L}_R(\mathbf{z}_\Gamma)$  and the *discrimination loss*  $\mathcal{L}_D(\mathbf{z}_\Gamma)$  at the last ( $\Gamma^{th}$ ) update iteration of the mapping procedure to the latent space, respectively. The model yields a large *anomaly score*  $A(\mathbf{x})$  for anomalous images, whereas a small *anomaly score* means that a very similar image was already seen during training. We use the *anomaly score*  $A(\mathbf{x})$  for image based anomaly detection. Additionally, the residual image  $\mathbf{x}_R = |\mathbf{x} - G(\mathbf{z}_\Gamma)|$  is used for the identification of anomalous regions within an image. For purposes of comparison, we additionally define a *reference anomaly score*  $\hat{A}(\mathbf{x}) = (1 - \lambda) \cdot R(\mathbf{x}) + \lambda \cdot \hat{D}(\mathbf{x})$ , where  $\hat{D}(\mathbf{x}) = \mathcal{L}_{\hat{D}}(\mathbf{z}_\Gamma)$  is the *reference discrimination score* used by Yeh et al. [13].

## 3 Experiments

*Data, Data Selection and Preprocessing* We evaluated the method on clinical high resolution SD-OCT volumes of the retina with 49 B-scans (representing an image slice in zx-plane) per volume and total volume resolutions of  $496 \times 512 \times 49$  voxels in z-, x-, and y direction, respectively. The GAN was trained on 2D image patches extracted from 270 clinical OCT volumes of healthy subjects, which were chosen based on the criterion that the OCT volumes do not contain fluid regions. For testing, patches were extracted from 10 additional healthy cases and 10 pathological cases, which contained retinal fluid. The OCT volumes were preprocessed in the following way. The gray values were normalized to range from -1 to 1. The volumes were resized in x-direction to a size of  $22\mu m$  resulting in approximately 256 columns. The retinal area was extracted and flattened to adjust for variations in orientation, shape and thickness. We used an automatic layer segmentation algorithm following [15] to find the top and bottom layer of the retina that define the border of the retina in z-direction. From these normalized and flattened volumes, we extracted in total 1.000.000 2D training patches with an image resolution of  $64 \times 64$  pixels at randomly sampled positions. Raw data and preprocessed image representation are shown in Figure 1. The test set in total consisted of 8192 image patches and comprised normal and pathological samples from cases not included in the training set. For pathological OCT scans, voxel-wise annotations of fluid and non-fluid regions from clinical retina experts were available. These annotations were only used for statistical evaluation but were never fed to the network, neither during training nor in the

evaluation phase. For the evaluation of the detection performance, we assigned a positive label to an image, if it contained at least a single pixel annotated as retinal fluid.

*Evaluation* The manifold of normal images was solely learned on image data of healthy cases with the aim to model the variety of healthy appearance. For performance evaluation in anomaly detection we ran the following experiments. (1) We explored qualitatively whether the model can generate realistic images. This assessment was performed on image patches of healthy cases extracted from the training set or test set and on images of diseased cases extracted from the test set.

(2) We evaluated quantitatively the anomaly detection accuracy of our approach on images extracted from the annotated test set. We based the anomaly detection on the *anomaly score*  $A(\mathbf{x})$  or only on one of both components, on the *residual score*  $R(\mathbf{x})$  or on the *discrimination score*  $D(\mathbf{x})$  and report receiver operating characteristic (ROC) curves of the corresponding anomaly detection performance on image level.

Based on our proposed *anomaly score*  $A(\mathbf{x})$ , we evaluated qualitatively the segmentation performance and if additional anomalies were identified.

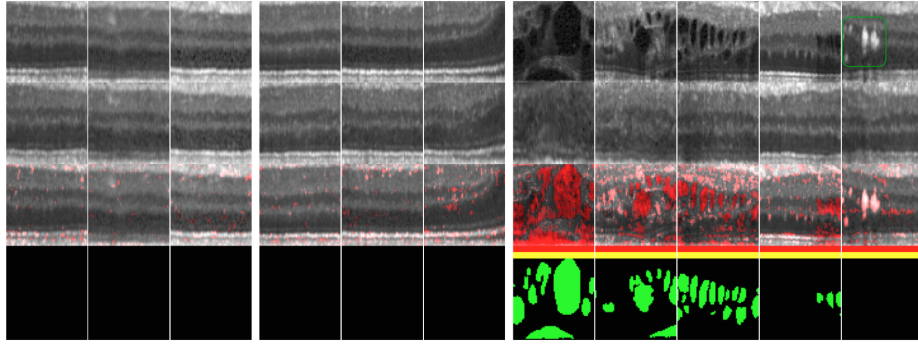
(3) To provide more details of individual components' roles, and the gain by the proposed approach, we evaluated the effect on the anomaly detection performance, when for manifold learning the adversarial training is not performed with a DCGAN but with an adversarial convolutional autoencoder (aCAE) [16], while leaving the definition of the *anomaly score* unchanged. An *aCAE* also implements a discriminator but replaces the generator by an encoder-decoder pipeline. The depth of the components of the trained *aCAE* was comparable to the depth of our adversarial model. As a second alternative approach, denoted as  $GAN_R$ , we evaluated the anomaly detection performance, when the *reference anomaly score*  $\hat{A}(\mathbf{x})$ , or the *reference discrimination score*  $\hat{D}(\mathbf{x})$  were utilized for anomaly scoring and the corresponding losses were used for the mapping from image space to latent space, while the pre-trained GAN parameters of the *AnoGAN* were used. We report ROC curves for both alternative approaches. Furthermore, we calculated sensitivity, specificity, precision, and recall at the optimal cut-off point on the ROC curves, identified through the Youden's index and report results for the *AnoGAN* and for both alternative approaches.

*Implementation details* As opposed to historical attempts, Radford et al. [12] identified a DCGAN architecture that resulted in stable GAN training on images of sizes  $64 \times 64$  pixels. Hence, we ran our experiments on image patches of the same size and used widely the same DCGAN architecture for GAN training (Section 2.1) as proposed by Radford et al. [12]<sup>1</sup>. We used four fractionally-strided convolution layers in the generator, and four convolution layers in the discriminator, all filters of sizes  $5 \times 5$ . Since we processed gray-scale images, we utilized intermediate representations with  $512 - 256 - 128 - 64$  channels (instead

---

<sup>1</sup> We adapted: <https://github.com/bamos/dcgan-completion.tensorflow>





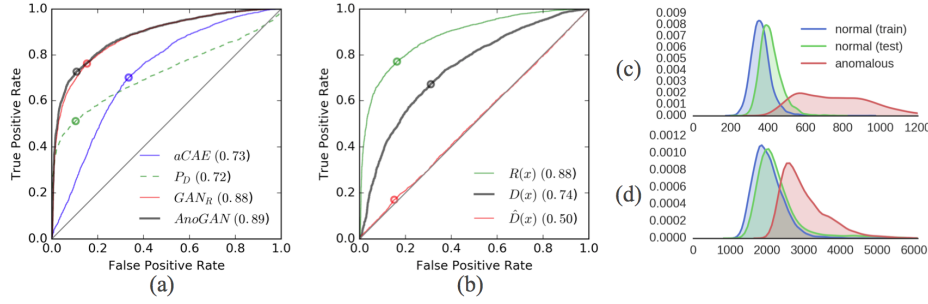
**Fig. 3.** Pixel-level identification of anomalies on exemplary images. First row: Real input images. Second row: Corresponding images generated by the model triggered by our proposed mapping approach. Third row: Residual overlay. Red bar: Anomaly identification by *residual score*. Yellow bar: Anomaly identification by *discrimination score*. Bottom row: Pixel-level annotations of retinal fluid. First block and second block: Normal images extracted from OCT volumes of healthy cases in the training set and test set, respectively. Third block: Images extracted from diseased cases in the test set. Last column: Hyperreflective foci (within green box). (Best viewed in color)

of  $1024 - 512 - 256 - 128$  used in [12]). DCGAN training was performed for 20 epochs utilizing Adam [17], a stochastic optimizer. We ran 500 backpropagation steps for the mapping (Section 2.2) of new images to the latent space. We used  $\lambda = 0.1$  in Equations (5) and (6), which was found empirically due to preceding experiments on a face detection dataset. All experiments were performed using Python 2.7 with the TensorFlow [18] library and run on a Titan X graphics processing unit using CUDA 8.0.

### 3.1 Results

Results demonstrate the generative capability of the DCGAN and the appropriateness of our proposed mapping and scoring approach for anomaly detection. We report qualitative and quantitative results on segmentation performance and detection performance of our approach, respectively.

**Can the model generate realistic images?** The trained model generates realistic looking medical images (second row in Figure 3) that are conditioned by sampling from latent representations  $\mathbf{z}$ , which are found through our mapping approach, described in Section 2.2. In the case of normal image patches (see first and second block in Figure 3), our model is able to generate images that are visually similar to the query images (first row in Figure 3). But in the case of anomalous images, the pairs of input images and generated images show obvious intensity or textural differences (see third block in Figure 3). The t-SNE embedding (Figure 2(b)) of normal and anomalous images in the feature representation of the last convolution layer of the discriminator that is utilized in the *discrimination loss*, illustrates the usability of the discriminator’s features for



**Fig. 4.** Image level anomaly detection performance and suitability evaluation. (a) Model comparison: ROC curves based on  $aCAE$  (blue),  $GAN_R$  (red), the proposed  $AnoGAN$  (black), or on the output  $P_D$  of the trained discriminator (green). (b) Anomaly score components: ROC curves based on the *residual score*  $R(\mathbf{x})$  (green), the *discrimination score*  $D(\mathbf{x})$  (black), or the *reference discrimination score*  $\hat{D}(\mathbf{x})$  (red). (c) Distribution of the *residual score* and (d) of the *discrimination score*, evaluated on normal images of the training set (blue) or test set (green), and on images extracted from diseased cases (red).

anomaly detection and suggests that our  $AnoGAN$  learns a meaningful manifold of normal anatomical variability.

**Can the model detect anomalies?** Figure 4(b) shows the ROC curves for image level anomaly detection based on the *anomaly score*  $A(\mathbf{x})$ , or on one of both components, on the *residual score*  $R(\mathbf{x})$ , or on the *discrimination score*  $D(\mathbf{x})$ . The corresponding area under the ROC curve (AUC) is specified in parentheses. In addition, the distributions of the *residual score*  $R(\mathbf{x})$  (Figure 4(c)) and of the *discrimination score*  $D(\mathbf{x})$  (Figure 4(d)) over normal images from the training set and test set or over images extracted from diseased cases show that both components of the proposed *adversarial score* are suitable for the classification of normal and anomalous samples. Figure 3 shows pixel-level identification of anomalies in conjunction with pixel-level annotations of retinal fluid, which demonstrate high accuracy. Last column in Figure 3 demonstrates that the model successfully identifies additional retinal lesions, which in this case correspond to hyperreflective foci (HRF). On image level, the red and yellow bars in Figure 3 demonstrate that our model successfully identifies every example image from diseased cases of the test set as being anomalous based on the *residual score* and the *discrimination score*, respectively.

**How does the model compare to other approaches?** We evaluated the anomaly detection performance of the  $GAN_R$ , the  $aCAE$  and the  $AnoGAN$  on image-level labels. The results are summarized in Table 1 and the corresponding ROC curves are shown in Figure 4(a). Although  $aCAEs$  simultaneously yield a generative model and a direct mapping to the latent space, which is advantageous in terms of runtimes during testing, this model showed worse performance on the anomaly detection task compared to the  $AnoGAN$ . It turned out that  $aCAEs$  tend to over-adapt on anomalous images. Figure 4(b) demonstrates that anomaly detection based on our proposed *discrimination score*  $D(\mathbf{x})$  outperforms

**Table 1.** Clinical performance statistics calculated at the Youden’s index of the ROC curve and the corresponding AUC based on the *adversarial score*  $A(\mathbf{x})$  of our model (*AnoGAN*) and of the *aCAE*, based on the *reference adversarial score*  $\hat{A}(\mathbf{x})$  utilized by  $GAN_R$ , or based directly on the output of the DCGAN ( $P_D$ ).

	Precision	Recall	Sensitivity	Specificity	AUC
aCAE	0.7005	0.7009	0.7011	0.6659	0.73
$P_D$	0.8471	0.5119	0.5124	0.8970	0.72
$GAN_R$	0.8482	0.7631	0.7634	0.8477	0.88
AnoGAN	0.8834	0.7277	0.7279	0.8928	0.89

the *reference discrimination score*  $\hat{D}(\mathbf{x})$ . Because the scores for the detection of anomalies are directly related to the losses for the mapping to latent space, these results give evidence that our proposed *discrimination loss*  $\mathcal{L}_D(\mathbf{z})$  is advantageous compared to the discrimination loss  $\mathcal{L}_{\hat{D}}(\mathbf{z})$ . Nevertheless, according to the AUC, computed based on the *anomaly score*, the *AnoGAN* and the  $GAN_R$  show comparable results (Figure 4(a)). This has to be attributed to the good performance of the *residual score*  $R(\mathbf{x})$ . A good anomaly detection performance (cf.  $P_D$  in Figure 4(a) and Table 1) can be obtained when the mapping to the latent space is skipped and a binary decision is derived from the discriminator output, conditioned directly on the query image.

## 4 Conclusion

We propose anomaly detection based on deep generative adversarial networks. By concurrently training a generative model and a discriminator, we enable the identification of anomalies on unseen data based on unsupervised training of a model on healthy data. Results show that our approach is able to detect different known anomalies, such as retinal fluid and HRF, which have never been seen during training. Therefore, the model is expected to be capable to discover novel anomalies. While quantitative evaluation based on a subset of anomaly classes is limited, since false positives do not take novel anomalies into account, results demonstrate good sensitivity and the capability to segment anomalies. Discovering anomalies at scale enables the mining of data for marker candidates subject to future verification. In contrast to prior work, we show that the utilization of the residual loss alone yields good results for the mapping from image to latent space, and a slight improvement of the results can be achieved with the proposed adaptations.

## References

1. Del Giorno, A., Bagnell, J.A., Hebert, M.: A discriminative framework for anomaly detection in large videos. In: ECCV, Springer (2016) 334–349
2. Matteoli, S., Diani, M., Theiler, J.: An overview of background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(6) (2014) 2317–2336

3. Carrera, D., Boracchi, G., Foi, A., Wohlberg, B.: Detecting anomalous structures by convolutional sparse models. In: 2015 International Joint Conference on Neural Networks (IJCNN), IEEE (2015) 1–8
4. Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition* **58** (2016) 121–134
5. Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. *Signal Processing* **99** (2014) 215–249
6. Venhuizen, F.G., van Ginneken, B., Bloemen, B., van Grinsven, M.J., Philipsen, R., Hoyng, C., Theelen, T., Sánchez, C.I.: Automated age-related macular degeneration classification in oct using unsupervised feature learning. In: SPIE Medical Imaging, International Society for Optics and Photonics (2015) 94141I–94141I
7. Schlegl, T., Waldstein, S.M., Vogl, W.D., Schmidt-Erfurth, U., Langs, G.: Predicting semantic descriptions from medical images with convolutional neural networks. In: International Conference on Information Processing in Medical Imaging. Volume 24., Springer (2015) 437–448
8. Seeböck, P., Waldstein, S., Klimesch, S., Gerendas, B.S., Donner, R., Schlegl, T., Schmidt-Erfurth, U., Langs, G.: Identifying and categorizing anomalies in retinal imaging data. NIPS 2016 MLHC workshop. preprint arXiv:1612.00686 (2016)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. (2014) 2672–2680
10. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in neural information processing systems. (2015) 1486–1494
11. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. arXiv:1605.09782 (2016)
12. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015)
13. Yeh, R., Chen, C., Lim, T.Y., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with perceptual and contextual losses. arXiv:1607.07539 (2016)
14. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems. (2016) 2226–2234
15. Garvin, M.K., Abramoff, M.D., Wu, X., Russell, S.R., Burns, T.L., Sonka, M.: Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *Transactions on Medical Imaging, IEEE* **28**(9) (2009) 1436–1447
16. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. *CoRR* **abs/1604.07379** (2016)
17. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
18. 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.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015) Software available from tensorflow.org.