SPARSE-GAN: SPARSITY-CONSTRAINED GENERATIVE ADVERSARIAL NETWORK FOR ANOMALY DETECTION IN RETINAL OCT IMAGE

Kang Zhou[1,2], *Shenghua Gao*[1,†], *Jun Cheng*[2,3,†], *Zaiwang Gu*[3], *Huazhu Fu*[5], *Zhi Tu*[1], *Jianlong Yang*[2], *Yitian Zhao*[2], *Jiang Liu*[2,4]

[1] School of Information Science and Technology, ShanghaiTech University
[2] Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences
[3] UBTech Research
[4] Southern University of Science and Technology
[5] Inception Institute of Artificial Intelligence

ABSTRACT

With the development of convolutional neural network, deep learning has shown its success for retinal disease detection from optical coherence tomography (OCT) images. However, deep learning often relies on large scale labelled data for training, which is oftentimes challenging especially for disease with low occurrence. Moreover, a deep learning system trained from data-set with one or a few diseases is unable to detect other unseen diseases, which limits the practical usage of the system in disease screening. To address the limitation, we propose a novel anomaly detection framework termed Sparsity-constrained Generative Adversarial Network (Sparse-GAN) for disease screening where only healthy data are available in the training set. The contributions of Sparse-GAN are two-folds: 1) The proposed Sparse-GAN predicts the anomalies in latent space rather than image-level; 2) Sparse-GAN is constrained by a novel Sparsity Regularization Net. Furthermore, in light of the role of lesions for disease screening, we present to leverage on an anomaly activation map to show the heatmap of lesions. We evaluate our proposed Sparse-GAN on a publicly available dataset, and the results show that the proposed method outperforms the state-of-the-art methods.

Index Terms— Anomaly Detection, Sparsity-constrained Network, Latent Feature, Adversarial Learning

1. INTRODUCTION

Over 300 million people worldwide are affected by various ocular diseases [1], such as diabetic retinopathy (DR) [2], age-related macular degeneration (AMD) and glaucoma. Among the many diagnostic methods, optical coherence tomography (OCT) is a non-invasive imaging modality that provides micrometer-resolution volumetric scans of the retina [3]. With the development of convolutional neural networks (CNNs) in computer vision [4, 5], many deep learning based approaches have been proposed to detect lesions in retinal OCT images [6] and fundus images [7, 8]. However, these deep learning based methods rely heavily on big data for training, which limits the application of deep learning to medical image analysis.

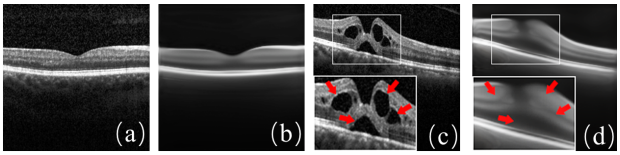

Fig. 1. The input image and its reconstructed image. (a) Normal input. (b) Reconstruction of the normal input. (c) Disease input. (d) Reconstruction of the disease input with our proposed method. Since lesions can't be reconstructed, the reconstruction error is high to be recognized as abnormal.

Different from that in the general computer vision, it is often challenging to get sufficient data for medical images due to several reasons. The first reason is that most of the medical data is not publicly available due to privacy concerns. The second reason is that labeling medical images often costs much time, while experienced clinicians are short of time for such tedious demarcation tasks. The third reason is that the occurrence of some lesions is usually low, while the presence of specific lesions is not known before the diagnosis. Therefore, the cost of obtaining large-scale medical data with particular types of lesions is often expensive and time-consuming.

Although it is difficult to get a large amount of data with different lesions, it is often much easier to get data from healthy subjects. In OCT imaging, one 3D scan from a healthy subject could provide hundreds of B-scan images without lesions. Considering the lesions as anomaly added to the images from healthy subjects, it is possible to train an anomaly detection system only using OCT B-scans without lesions.

[1] {zhoukang, gaoshh}@shanghaitech.edu.cn, [2] chengjun@nimte.ac.cn
[†] corresponding authors

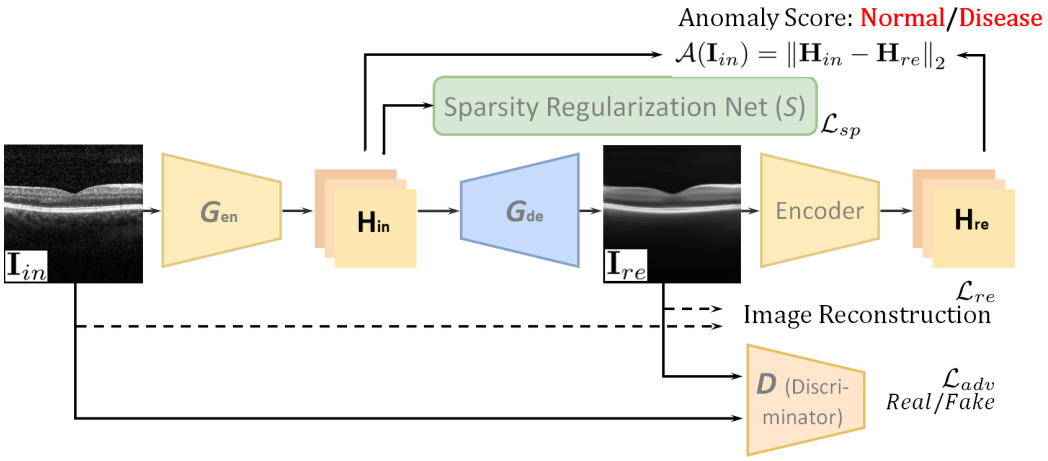

Fig. 2. The overall architecture of our Sparse-GAN. Components with boxes with solid line are networks while other boxes are features. In the **testing stage**, given a test image $\mathbf{I}_{in}$, firstly the image is converted into latent feature with $\mathbf{H}_{in} = G_{en}(\mathbf{I}_{in})$, while $\mathbf{H}_{in}$ is converted into reconstructed image with $\mathbf{I}_{re} = G_{de}(\mathbf{H}_{in})$. Then $\mathbf{I}_{re}$ is transformed to latent feature with another encoder $\mathbf{H}_{re} = E(\mathbf{I}_{re})$, finally the framework predicts anomaly score with $\mathcal{A}(\mathbf{I}_{in}) = \|\mathbf{H}_{in} - \mathbf{H}_{re}\|_2$. In the **training stage**, besides the same pipeline of testing, the framework is trained with image reconstruction loss $\mathcal{L}_{re}$, adversarial loss $\mathcal{L}_{adv}$ and sparsity regularization $\mathcal{L}_{sp} = S(\mathbf{H}_{in})$. (Best viewed with colors.)

Previous work has shown the effectiveness of anomaly detection for disease diagnosis [9] and lesion location [10]. Recently, CNNs based methods have been proposed to detect anomalies in medical images. Schlegl *et al.* [11] initially introduce a deep convolutional Generative Adversarial Network (GAN) [12], which is referred to a AnoGAN, to detect anomalies in OCT B-scans. Later, they further propose a f-AnoGAN [13], which is faster than AnoGAN. However, these networks are not trained in an end-to-end fashion, which may tend to get stuck into local optima. It is desirable to customize a network that learns the optimal features for anomaly detection.

In this paper, inspired by Image-to-Image GAN [14], whose generator is end-to-end optimized, we propose to employ Image-to-Image GAN for medical image anomaly detection. Then, to alleviate the effect of image noise (e.g. speckle noise in OCT images), we propose to map the reconstructed image into latent space with an additional encoder. Furthermore, motivated by the capability of interpretable sparse coding for anomaly detection, we propose to regularize the sparsity of latent features. By taking these factors into consideration, we present a novel framework: Sparsity-constrained Generative Adversarial Network (Sparse-GAN) for image anomaly detection with merely normal training data. The rationale behind the work is that the normal patterns from healthy subjects can be reconstructed with small errors while the patterns with lesions from diseased subjects are often reconstructed with large errors, as shown in Fig. 1.

The main contributions of this work are summarized as follows: (1) We propose to map the images into a latent space and regularize the latent feature with a novel sparsity regularizer; (2) We introduce a novel Sparse-GAN for anomaly detection, and our method is designed for the scenario where only data corresponding to healthy subjects are available in the training set. Thus our solution may ease the difficulty in data collection and annotation; (3) Our method also predicts anomaly activation maps to show lesions for clinical diagnosis.

2. METHOD

In this work, we mainly focus on regularizing the sparsity of latent feature and utilizing the latent feature to predict anomalies in GAN based anomaly detection framework. As shown in Fig. 2, the proposed Sparsity-constrained Generative Adversarial Network consists of three modules: 1) Image-to-Image GAN [14] for medical anomaly detection whose generator is end-to-end optimized. 2) Anomalies computing in latent space [15], to alleviate the effect of image noise (e.g. speckle noise in OCT images). 3) The novel Sparsity Regularization Net to regularize the sparsity of latent features.

2.1. Image-to-Image GAN for Anomaly Detection

As discussed earlier, we adopt the image-to-image [14] generator as the G in the GAN, which consists of encoder G_{en} and decoder G_{de}, while D denotes the discriminator. Let $\mathbf{I}_{in}$ be input images, their latent feature $\mathbf{H}_{in}$ are converted from input images $\mathbf{H}_{in} = G_{en}(\mathbf{I}_{in})$, then the latent feature are transformed into reconstructed images $\mathbf{I}_{re} = G_{de}(\mathbf{H}_{in})$. Image-to-Image GAN [14] is optimized with a reconstruction loss comprised of an adversarial loss,

$$\min_G \max_D \mathcal{L}_G = \min_G \left(\lambda_{adv} \max_D (\mathcal{L}_{adv}) + \lambda_{re} \mathcal{L}_{re} \right), \quad (1)$$

where λ_{adv} and λ_{re} are regularization parameters. The adversarial loss and reconstruction loss are defined as,

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{I}_{in}}[\log D(\mathbf{I}_{in})] + \mathbb{E}_{\mathbf{I}_{in},\mathbf{H}_{in}}[\log(1 - D(G(\mathbf{I}_{in}), \mathbf{H}_{in}))], \tag{2}$$

$$\mathcal{L}_{re} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{I}_{in}^{(i)} - \mathbf{I}_{re}^{(i)})^2, \tag{3}$$

where m is the batch-size.

2.2. Predict Anomaly Score in Latent Space

One challenge in reconstructing the OCT images is the speckle noise. To reduce the influence of speckle noise, we propose to transform the reconstruction image $\mathbf{I}_{re}$ into latent space by encoder E, i.e. $\mathbf{H}_{re} = E(\mathbf{I}_{re})$. To cut down computational cost, encoder E share the same values with G_{en}. In latent space, the model predicts anomaly score $\mathcal{A}(\mathbf{I}_{in})$ and diagnosis results $\mathcal{C}(\mathbf{I}_{in})$ as follows:

$$\mathcal{A}(\mathbf{I}_{in}) = \|\mathbf{H}_{in} - \mathbf{H}_{re}\|_2 = \|G_{en}(\mathbf{I}_{in}) - E(G(\mathbf{I}_{in}))\|_2, \tag{4}$$

$$\text{and } \mathcal{C}(\mathbf{I}_{in}) = \begin{cases} \text{normal}, & \text{if } \mathcal{A}(\mathbf{I}_{in}) < \phi \\ \text{disease}, & \text{if } \mathcal{A}(\mathbf{I}_{in}) \geqslant \phi \end{cases} \tag{5}$$

where ϕ is the anomaly score threshold determined on the validation set.

2.3. Sparse Regularization on Latent Feature

On the one hand, without additional regularization, generator G may learn an approximation to the identity function, which can not distinguish disease images from normal images. On the other hand, sparse coding is interpretable and have the capability for anomaly detection [16, 17].

Based on this observation, we propose a novel Sparsity Regularization Net which recast the solution of sparse coding as a novel convolutional long short term memory unit (LSTM). Moreover, we regularize the sparsity of latent feature $\mathbf{H}_{in}$ with the proposed Sparsity Regularization Net (i.e., $S(\cdot)$) as shown in Fig. 2. Letting S denote Sparsity Regularization Net, we propose a novel Sparsity-constrained GAN (Sparse-GAN) with sparsity regularization $\mathcal{L}_{sp} = S(\mathbf{H}_{in})$.

The proposed Sparsity Regularization Net is inspired from Sparse LSTM [18]. However, sparsity reguliarzaiton net is different from sparse LSTM in two aspects. Firstly we apply the convolutional operation to replace element-wise multiplication in Sparse LSTM since the convolutional operation accelerates the computation. Secondly the input of the Sparse Constrained Net is the latent feature rather than the original image.

The loss to train Sparsity Regularization Net is defined as follows,

$$\mathcal{L}_{scl}(\mathbf{W}_d, \mathbf{s}) = \left\|\mathbf{H}_{in} - \mathbf{W}_d^T\mathbf{s}\right\|_F^2 + \|\mathbf{s}\|_1 \tag{6}$$

where $\mathbf{s}$ is the sparse code w.r.t. $\mathbf{H}_{in}$ and $\mathbf{W}_d$ is the dictionary.

Overall, the final loss of Sparse-GAN is given as the following:

$$\mathcal{L} = \lambda_{re}\mathcal{L}_{re} + \lambda_{adv}\max_{D}(\mathcal{L}_{adv}) + \lambda_{sp}\mathcal{L}_{sp}, \tag{7}$$

where $\lambda_{re}, \lambda_{adv}$ and λ_{sp} are regularization parameters.

2.4. Anomaly Activation Map for Visualization

Since anomaly detection is significantly different from supervised classification, Class Activation Map (CAM) [19] is not suitable in our framework to show the role of lesions for diagnosis. To address the weakness of CAM, we propose Anomaly Activation Map (AAM) to visualize lesions in anomaly detection framework. We firstly perform Global Average Pooling (GAP) for latent feature $\mathbf{H}_{in} \in \mathbb{R}^{1024\times7\times7}$ and $\mathbf{H}_{re} \in \mathbb{R}^{1024\times7\times7}$. Then we obtain the anomaly vector $\mathbf{W}_{aam} = w_1, w_2, \cdots, w_n$ as follows,

$$\mathbf{W}_{aam} = \|GAP(\mathbf{H}_{in}) - GAP(\mathbf{H}_{re})\|_1, \tag{8}$$

where $\mathbf{W}_{aam} \in \mathbb{R}^{1024\times1\times1}$, n is the number of the channels of the latent feature. Finally, we multiply the feature map $\mathbf{H}_{in}$ by anomaly vector in channel-wise fashion and get the anomaly activation map.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

3.1.1. Datasets

We employ a publicly available dataset [20] to evaluate the performance of our Sparse-GAN. The whole dataset was from Spectralis OCT (Heidelberg Engineering, German), and contains data with three different lesions: drusen, DME (diabetic macular edema), and CNV (choroidal neovascularization). The detailed description about this dataset could be found in [20]. To train the proposed Sparse-GAN and determine the threshold of anomaly score, we divide original training set into two parts: new training set with 50,140 normal images, validation set consists of 3000 disease images and 1000 normal images. The testing set is the same as the original dataset.

3.1.2. Evaluation Metrics

For a given test image $\mathbf{I}_{in}$, we use $\mathcal{A}(\mathbf{I}_{in})$ given in Eq. (4) to compute the anomaly score. Further, we use $\mathcal{C}(\mathbf{I}_{in})$ given in Eq. (5) for diagnosis. Based on the anomaly score, we mainly use **AUC** (Area under the ROC Curve) to evaluate our method. To compute accuracy (**Acc**), we need to determine the threshold ϕ of anomaly score on the validation set, which includes 75% disease images and 25% normal images. We adopt sensitivity (**Sen**) as the third evaluation metric. Finally, the threshold ϕ is then used for testing.

3.2. Training Details

The proposed Sparse-GAN is implemented in PyTorch with NVIDIA graphics processing units (GeForce TITAN V). The input image size is 224×224, while the batch size is 32. The optimizer is Adam and the learning rate is 0.001. Empirically, we let $\lambda_{re} = 20$, $\lambda_{adv} = 1$, and $\lambda_{sp} = 50$.

3.3. Quantitative Experimental Results

Table 1. Quantitative results for ablation studies and comparison with state-of-the-arts.

Method	Val-set	Test-set		
	AUC	AUC	Acc	Sen
Auto-Encoder	0.729	0.783	0.751	0.834
AnoGAN[11]	0.815	0.846	0.789	0.917
f-AnoGAN[13]	0.849	0.882	0.808	0.871
pix2pix [14] #1	0.805	0.861	0.818	0.879
pix2pix [14] #2	0.837	0.874	0.815	0.900
Sparse-GAN	**0.885**	**0.925**	**0.841**	**0.951**

#1, image level
#2, latent space

3.3.1. Ablation Study.

To justify the benefits of the anomaly score in latent space and the sparsity regulirization nets, we conduct the following ablation studies, we conduct some ablation studies: #1 denotes Image-to-Image GAN [14] predicting anomaly score in image-level, and #2 denotes Image-to-Image GAN [14] predicting anomaly score $\mathcal{A}(\mathbf{I}_{in})$ in latent feature.

By including $\mathcal{L}_{adv}$ loss based on Auto-Encoder, we improve the AUC result from 0.729 to 0.805 on the validation set. That is to say, adversarial learning is helpful. By transforming the reconstruction image into latent space, the result is improved from 0.805 to 0.837 on the validation set since the noise in images is harmful to diagnosis. Finally, by regularizing the latent features with our proposed Sparsity Regularization Net, the result is improved from 0.837 to 0.885, which means the sparsity regularization is effective. On the test set, the ablation studies validate the effectiveness of different modules too. Table 1 summarized the results.

3.3.2. Performance Comparison.

We further compare the proposed method with state-of-the-art networks, inlcuding Auto-Encoder, AnoGAN [11] and f-AnoGAN [13].

By comparing our adopted Image-to-Image GAN (i.e. # 1) with primary AnoGAN [11], we improve the AUC result from 0.846 to 0.861 on the test set. That is to say, the end-to-end optimized generator is better than two stage trained generator. Compared with these methods, we get the highest AUC than others on both the validation set and test set. The accuracy of our method on the test set is comparable to supervised deep learning methods, and the sensitivity $= 0.951$ denotes missed diagnosis of our model is very low, which is more meaningful for clinicians. The results are also summarized in Table 1 .

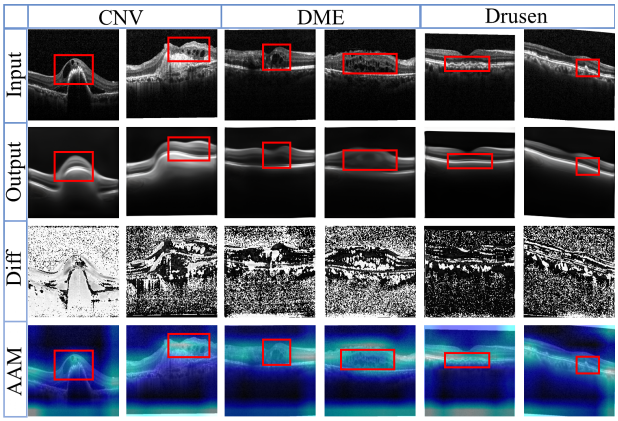

Fig. 3. Anomaly heatmap on abnormal images. *Diff* images show that noise in images is harmful for reconstruction, and AAM images show the lesion play an important role for diagnosis in Sparse-GAN. (Best viewed with colors.)

3.4. Qualitative Analysis with Anomaly Activation Map

To further understand what the role of the lesion is for disease clinical diagnosis, some example images are shown in Fig 3. When Sparse-GAN classifies a given image as abnormal, AAM will be computed. In addition to the anomaly heatmap, we also show the output images and difference between the input image and output one. Since Sparse-GAN is only trained on the normal set, the model could not reconstruct abnormal patterns. *Diff* images show that noise in images is harmful to reconstruction. The heatmap can localize the lesion in general and this validates the effectiveness of our proposed AAM for anomaly detection framwork.

4. CONCLUSION

In this work, we propose a novel Sparse-GAN for anomaly detection, which detects anomalies in latent space and the feature in latent space is constrained by a novel Sparsity Regularizer Net. The quantitative experimental results on a public dataset validate the feasibility of anomaly detection for OCT images and also validate the effectiveness of our method. Further, we also show the anomaly activation maps of the lesion to make our results more explainable.

5. ACKNOWLEDGE

The project is partially supported by ShanghaiTech-Megavii Joint Lab, in part by the National Natural Science Foundation of China (NSFC) under Grants No. 61932020, and supported by the ShanghaiTech-UnitedImaging Joint Lab, Ningbo "2025 S&T Megaprojects" and Ningbo 3315 Innovation team grant. We also acknowledge the contribution of Weixin Luo and Wen Liu for their insightful comments with regard to the reconstruction-based anomaly detection method.

6. REFERENCES

[1] Stefanos Apostolopoulos, Sandro De Zanet, et al., "Pathological oct retinal layer segmentation using branch residual u-shape networks," in *MICCAI*. Springer, 2017, pp. 294–301.

[2] Yitian Zhao, Yalin Zheng, et al., "Uniqueness-driven saliency analysis for automated lesion detection with applications to retinal diseases," in *MICCAI*. Springer, 2018, pp. 109–118.

[3] David Huang, Eric A Swanson, et al., "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[4] Alex Krizhevsky, Ilya Sutskever, et al., "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[5] Dongze Lian, Lina Hu, et al., "Multiview multitask gaze estimation with deep convolutional neural networks," *IEEE transactions on neural networks and learning systems*, 2018.

[6] Cecilia S Lee, Doug M Baughman, et al., "Deep learning is effective for classifying normal versus age-related macular degeneration oct images," *Ophthalmology Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[7] Kang Zhou, Zaiwang Gu, et al., "Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2018, pp. 2724–2727.

[8] Zaiwang Gu, Jun Cheng, et al., "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, 2019.

[9] Desire Sidibe, Shrinivasan Sankar, et al., "An anomaly detection approach for the identification of dme patients using spectral domain optical coherence tomography images," *Computer methods and programs in biomedicine*, vol. 139, pp. 109–117, 2017.

[10] Philipp Seeböck, Sebastian M Waldstein, et al., "Unsupervised identification of disease marker candidates in retinal oct imaging data," *IEEE TMI*, 2018.

[11] Thomas Schlegl, Philipp Seeböck, et al., "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *IPMI*. Springer, 2017, pp. 146–157.

[12] Ian Goodfellow, Jean Pouget-Abadie, et al., "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[13] Thomas Schlegl, Philipp Seeböck, et al., "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, 2019.

[14] Phillip Isola, Jun-Yan Zhu, et al., "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[15] Samet Akcay, Amir Atapour-Abarghouei, et al., "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 622–637.

[16] Weixin Luo, Wen Liu, et al., "A revisit of sparse coding based anomaly detection in stacked rnn framework," *ICCV, Oct*, vol. 1, no. 2, pp. 3, 2017.

[17] Weixin Luo, Wen Liu, et al., "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[18] Joey Tianyi Zhou, Kai Di, et al., "Sc2net: Sparse lstms for sparse coding," in *AAAI*, 2018.

[19] Bolei Zhou, Aditya Khosla, et al., "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[20] Daniel S Kermany, Michael Goldbaum, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.