BR-GAN: Bilateral Residual Generating Adversarial Network for Mammogram Classification

Chu-ran Wang[1,2,5], Fandong Zhang[1(✉)], Yizhou Yu[5], and Yizhou Wang[2,3,4]

[1] Center for Data Science, Peking University, Beijing, China
fd.zhang@pku.edu.cn
[2] Advanced Institute of Information Technology, Peking University, Hangzhou, China
[3] Center on Frontiers of Computing Studies, Peking University, Beijing, China
[4] Department of Computer Science, Peking University, Beijing, China
[5] Deepwise AI Lab, Beijing, China

Abstract. Mammogram malignancy classification with only image-level annotations is challenging due to a lack of lesion annotations. If we can generate the healthy version of the diseased data, we can easily explore the lesion features. An intuitive idea of such generation is to use existing Cycle-GAN based methods. They achieve the healthy generation regarding healthy images as reference domain, while maintaining the original content by cycle consistency mechanism. However, healthy mammogram patterns are diverse which may lead to uncertain generations. Moreover, the back translation from healthy to the original remains an ill-posed problem due to lack of lesion information. To address these problems, we propose a novel model called bilateral residual generating adversarial network(BR-GAN). We use the Cycle-GAN as a basic framework while regarding the contralateral as generation reference based on the bilateral symmetry prior. To address the ill-posed back translation problem, we propose a residual-preserved mechanism to try to preserve the lesion features from the original features. The generated features and the original features are aggregated for further classification. BR-GAN outperforms current state-of-the-art methods on INBreast and in-house datasets.

Keywords: Mammogram classification · Domain knowledge · Cycle consistency mechanism

1 Introduction

Breast cancer is the most commonly diagnosed cancer among women [15]. Mammography is a common examination for early breast cancer diagnosis. The mammogram malignancy classification is crucial. Most existing methods require extra annotations, such as bounding boxes for detection [1,8,12,16,17] and mask ground truth for segmentation [7]. However, the above extra annotations require

C. Wang—This work was done when Chu-ran Wang was an intern at Deepwise AI Lab.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12262, pp. 657–666, 2020.
https://doi.org/10.1007/978-3-030-59713-9_63

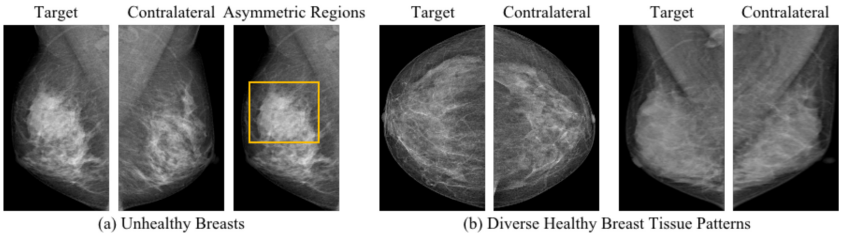

Fig. 1. Cases to show that how the unhealthy breasts look asymmetrical, while healthy breasts are roughly symmetrical

expert domain knowledge, which is costly to obtain. Therefore, mammogram malignancy classification with the only image-level labels as the supervision is of vital significance clinically.

Exploring lesion features from a full mammogram image is the key to solve the problem. However, lesion exploration is very challenging since the lesions can be expressed as diverse appearances and the high-intensity breast tissues may partially obscure the lesions. Previous researches mainly use attention mechanism for abnormal exploration, *e.g.*, Zhu *et al.* [20] and Fukui *et al.* [2]. However, the lack of using mammogram domain knowledge limits their performances.

Learning a healthy generation could be an effective way to exploit domain structure prior. Given a diseased image, if we know how its healthy version behaves we can localize the abnormal regions easily by the difference between the original and its healthy version. Thus such prior provides a more direct and credible way to localize abnormalities from a full mammogram. AnoGAN [13] applies such thought to anomaly detection. However, training with only healthy images restricts its effectiveness in our application. Fixed-Point GAN [14] and CycleGAN [19] can be used for the healthy generation based on the cycle consistency mechanism. An intuitive idea of applying to our application is to regard unhealthy images as a domain and healthy images as another domain. However, such approaches have two major limitations. First, we need to know which images are healthy. Moreover, healthy patterns in mammograms can be various and even similar to the lesions in some cases as shown in Fig. 1. We want to generate a healthy image that maintains all the healthy contents of the original. Regarding healthy images as generation reference may lead to a diverse generation and may conflict with our goal. Second, the cycle consistency mechanism assumes that the translated data can be translated back to the original data [5,10] and leads to the preservation of the original features, *e.g.*, large objects and textures. However, lesions in our application can appear anywhere and have diverse appearances. It translates the healthy domain back to the original domain an ill-posed task. Thus such methods will result in undesirable lesion removal in our application.

To address the first problem, we directly regard the contralateral as generation reference by making use of the mammogram bilateral symmetry prior. To be clear, we call the image to be classified as the target, while the image of the opposite side as the contralateral. Bilateral breasts from the same person have

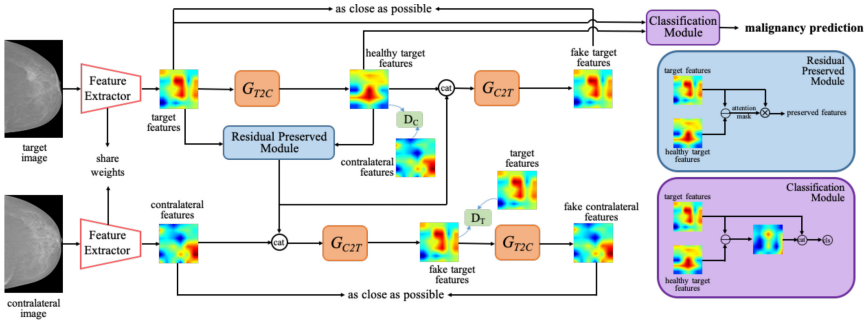

Fig. 2. The schematic overview of BR-GAN. Framework is based on CycleGAN [19] but uses contralateral features as references and adds preserved features calculated in *Residual-Preserved Module* to the back translation network. The generated features are then fed into a *Classification Module* with target features. Finally the *Classification Module* outputs labels of benign/malignant.

a roughly symmetrical glandular pattern. Most lesions only appear on one side and are invisible in the symmetrical regions of the opposite side. Therefore, the contralateral can be an effective reference for generating the healthy version of the target. Besides, a standard mammogram contains images from both sides. Thus, no extra data is required. To tackle the second problem, *i.e.,* ill-posed back translation, we try to preserve the suspicious lesion information while translating from the target data to its healthy version, and plus it when translating back. Thus, the information fed to the back translation is supposed to be sufficient.

In this paper, we propose a novel model named Bilateral Residual Generating Adversarial Network(BR-GAN) to improve mammogram malignancy classification by making use of bilateral prior and healthy generation mechanism. First, we propose a bilateral-cycle mechanism. We use contralateral images as references instead of healthy images. Due to the bilateral misalignment problem, we perform the generation in feature-level. Second, we propose a residual-preserved mechanism for better preserving lesion information during translation. While generating healthy features, we preserve the target-healthy residual features with the attention mechanism. We constrain the preserved features and the target features to share the same malignancy prediction by a residual embedding loss. In the healthy/contralateral-target translation, the preserved features are also fed into the translation network. Finally, we aggregate the generated features with the target features for further classification. Experimental results on both the public dataset and the in-house dataset demonstrate the proposed BR-GAN achieves state-of-the-art performance.

2 Bilateral Residual Generating Adversarial Network

Figure 2 outlines the overall network architecture of our framework. We first use contralateral features as references and generate the healthy version of the

target features (Sect. 2.1). Then we feed both the target features and the residual between the generated features and the target features into *Classification Module* to predict labels (Sect. 2.2). We design our model in feature-level instead of pixel-level due to the bilateral misalignment.

2.1 Feature Generation

Generated features are the healthy version of the target features. To achieve feature generation, we propose a bilateral-cycle mechanism based on a Cycle-GAN framework. Due to the limitation of the cycle mechanism, we design a residual-preserved mechanism to provide lesion information for translation from healthy to unhealthy.

Bilateral-Cycle Mechanism. The GAN loss can be defined as Eq. 1.

$$\begin{aligned}\min_{G}\max_{D}\mathcal{L}_G(G,D,f_{target},f_{reference}) := &\log\left(D\left(f_{reference}\right)\right)\\ &+\log\left(1-D\left(G(f_{target})\right)\right),\end{aligned} \tag{1}$$

where $f_{reference}$ is defined as the features of the reference used in discriminator D and f_{target} is defined as the features of the target image which needs to be classified.

Most paired healthy breasts are roughly symmetrical and the abnormalities are rarely symmetrical. Thus, contralateral features are appropriate references for the healthy generation. We use contralateral features as references in our basic Cycle-GAN framework. Generator G_{T2C} tries to generate healthy features f_H^T that look similar to the contralateral features f^C from the target features f^T, while D_C aims to distinguish between translated features f_H^T and real features f^C. Generator G_{T2C} is optimized by $\underset{G_{T2C}}{min}\underset{D_C}{max}L_{G_{T2C}}(G_{T2C}, D_C, f^T, f^C)$, and $f_H^T = G_{T2C}(f^T)$. While generator G_{C2T} tries to translate the generated healthy features f_H^T back to the target features f^T and help f_H^T maintain the target features in lesion-free areas.

However, lesions in our application can appear anywhere and have multiple shapes. Due to the limitation of the cycle consistency mechanism mentioned in Sect. 1, the generated healthy features can not provide lesion information for back translation. Thus it will be an ill-posed problem if we feed the generated features f_H^T into the generator G_{C2T} directly. We propose a residual preserved mechanism to tackle the problem.

Residual Preserved Mechanism. While we translate the target features to its healthy version, we separate the suspicious lesion features *i.e.,* the preserved features f_P^T from the target features in *Residual Preserved Module*. The preserved features f_P^T are used as the guidance to indicate the predicted lesion information. Thus the preserved features f_P^T should contain the texture and space information of the lesions. Concatenation of the preserved features f_P^T and the generated features f_H^T will be inputs to the generator G_{C2T}, *i.e.* $G_{C2T}([f_H^T, f_P^T])$. The preserved features f_P^T will provide lesion features for a back translation and avoid the ill-posed problem.

To calculate the preserved features, we first calculate the residual between the target features and the generated features. Second, to avoid the back translation network G_{C2T} being a direct identity mapping, we do not use residual as our preserved features directly. We turn the residual features into an attention map by softmax function for normalization. If the generated features learn to be the healthy version of the target features, the locations on residual features with high values should indicate high abnormal probabilities. Third, we multiply the attention map and the target features. Finally, we get the preserved features f_P^T which is defined as:

$$f_P^T = f^T * softmax(f^T - f_H^T) \tag{2}$$

To further constrain the success of separation, we define a residual embedding loss Eq. 3 and constrain the preserved features and the target features to share the same malignancy prediction. We use the malignancy classifier to predict the malignant probabilities $p_m(\cdot)$ of the target features f^T and the preserved features f_P^T.

$$\mathcal{L}_{RE} = -p_m(f^T) * log(p_m(f_P^T)) - (1 - p_m(f^T)) * log(1 - p_m(f_P^T)). \tag{3}$$

We design the residual cycle consistency loss L_c^T measured by selected mean square error(MSE) to achieve $f^T \rightarrow G_{T2C}(f^T) \rightarrow G_{C2T}([G_{T2C}(f^T), f_P^T]) \approx f^T$.

However, with only one residual cycle consistency constrain may lead to a collapsed identical mapping from contralateral features. To avoid this problem, we design another cycle consistency loss L_c^C. L_c^C also is measured by MSE and achieves *i.e.,* $f^C \rightarrow G_{C2T}([f^C, f_P^T]) \rightarrow G_{T2C}(G_{C2T}([f^C, f_P^T])) \approx f^C$. And the generator G_{C2T} is optimized by $\underset{G_{C2T}}{min}\underset{D_T}{max} L_{G_{C2T}}(G_{C2T}, D_T, f^C, f^T)$, while D_T aims to distinguish between translated features $G_{C2T}([f^C, f_P^T])$ and real target features f^T.

2.2 Classification

From the feature generation procedure, we obtain the healthy features f_H^T of the target image x^T. The healthy features f_H^T and the target features f^T are fed into *Classification Module* for final classification. In the module, we first calculate the residual between the generated features f_H^T and the original target features f^T. Then concatenation of the residual and the original target features f^T which contain global semantic information is used to predict benign-malignant labels. We use the cross-entropy loss as loss function L_{CLS} for mammogram classification.

During training, we optimize both feature generation and classification modules jointly as in Eq. 4.

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{RE} + \mathcal{L}_{CLS} + \min_{G_{T2C}} \max_{D_C} \mathcal{L}_{G_{T2C}}(G_{T2C}, D_C, f^T, f^C) \\ + \min_{G_{C2T}} \max_{D_T} \mathcal{L}_{G_{C2T}}(G_{C2T}, D_T, f^C, f^T) + \mathcal{L}_c^T + \mathcal{L}_c^C\end{aligned} \tag{4}$$

3 Experiments

3.1 Experimental Settings

Datasets. We evaluate BR-GAN on a public INBreast dataset [9] and an in-house dataset. INBreast [9] has 115 cases and 410 mammograms and provides each image a BI-RADS result as image-wise ground truth. We use the same process as Zhu *et al.* [20] (malignant if BI-RADS > 3; benign otherwise). For a fair comparison, our settings are all the same as Zhu *et al.* [20] for mass classification on the INBreast [9]. However, we discard 9 images for the lack of contralateral images and the remainings all have contralateral images. In addition, we also attempt mixed-lesion classification including mass, calcification cluster and distortion for the purpose of generalization. The in-house dataset contains 1303 images with malignancy annotations, including 589 only masses,120 only suspicious calcifications,34 only architectural distortions, 197 only asymmetries and 363 multiple lesions from 642 patients. All these 1303 images have opposite sides, i.e. 1303 pairs. We randomly divide the dataset into training, validation and testing sets as 8:1:1 in patient-wise.

Implementation Details. We use Otsus method [11] to segment the breast regions and remove backgrounds from the original images in 14-bit DICOM format. We implement all models with PyTorch and use Adam optimization. Both target and contralateral features are extracted from the last convolution layer. We use Area Under the Curve (AUC) as evaluation metrics in image-wise.

3.2 Performances

Mass Classification. The first four lines in Table 1 summarize the results of the representative methods. To be fair, we compare the results with the backbone of AlexNet [6] and ResNet50 [4] separately. Due to the slight difference of images caused by reference absence, for a fair comparison, we re-implement some representative methods for mammogram classification [20], natural image classification [2,18] and healthy generation [13,14,19] by adjusting the source codes given by the authors. We marked these methods by '*' in the table.

Mix-Lesion Classification. The performances are shown in the last two columns of Table 1 for the INBreast dataset and the in-house dataset.

Results. Attention mechanism (Zhu [20], CAM [18], ABN [2]) works but limits by the lack of mammogram domain knowledge. Only using healthy data for training highly (AnoGAN [13]) relies on the number of healthy data and is limited by the lack of reference to unhealthy data. The cycle consistency mechanism (Fixed-Point GAN [14], CycleGAN [19]) is effective to some extent but is limited by its ill-posed back translation problem in our application. However, our BR-GAN outperforms the representative methods significantly on both datasets.

To further evaluate the effectiveness of the generated features(healthy version of the target features), we calculate the mean FID [3] to measure the average of features distribution distances in the INBreast dataset. The mean FID between

Table 1. AUC evaluation on (a) INBreast for mass classification with Alexnet; (b) INBreast for mass classification with Resnet50; (c) INBreast for mixed-lesion classification with Resnet50; (d) in-house dataset for mixed-lesion classification with Alexnet.

Methodology	AUC (a)	AUC (b)	AUC (c)	AUC (d)
Pretrained CNN [1]	0.690	–	–	–
Pretrained CNN+RF [1]	0.760	–	–	–
Vanilla AlexNet, Zhu *et al.* [20]	0.790	–	–	–
Zhu *et al.* [20]	0.890	–	–	–
Vanilla*	0.820	0.827	0.780	0.697
AnoGAN [13]*	0.803	0.796	0.774	0.720
Fixed-Point GAN [14]*	0.835	0.837	0.805	0.734
CycleGAN [19]*	0.852	0.838	0.808	0.741
Zhu *et al.* [20]*	0.860	0.862	0.830	0.720
Vanilla*+GAP [18]*	0.857	0.827	0.780	0.718
Vanilla*+ABN [2]*	0.858	0.846	0.814	0.723
Proposed Method	**0.900**	**0.886**	**0.860**	**0.770**

Table 2. Top-1 localization error on (b) INBreast dataset for mass classification with Resnet50; (d) INBreast dataset for mixed-lesion classification with Resnet50.

Methodology	Top-1 error (b)	Top-1 error (d)
ResNet50 [4]	0.635	0.727
AnoGAN [13]*	0.684	0.789
Fixed-Point GAN [14]*	0.646	0.737
CycleGAN [19]*	0.632	0.667
ABN [2]	0.632	0.722
Zhu *et al.* [20]*	0.627	0.625
Proposed Method	**0.519**	**0.544**

the target and contralateral features is 63.63. The generated-contralateral mean FID is 27.54. The target-generated mean FID is 22.81 while the one after removing the lesion areas from ground truth is 0.73. Through the above comparison, we can find the generated features containing both contralateral distribution and target information in healthy areas as we want.

Localization. To verify whether the proposed model focuses on the lesion areas, we evaluate the localization error by CAM [18]. We use the top-1 localization error as ILSVRC using an inter-over-union (IOU) threshold of 0.1. As is shown in Table 2, BR-GAN largely outperforms the representative methods.

Furthermore, Fig. 3 visualizes the class activation maps of some cases. As we can see, all lesions satisfy the bilateral asymmetry prior. The proposed

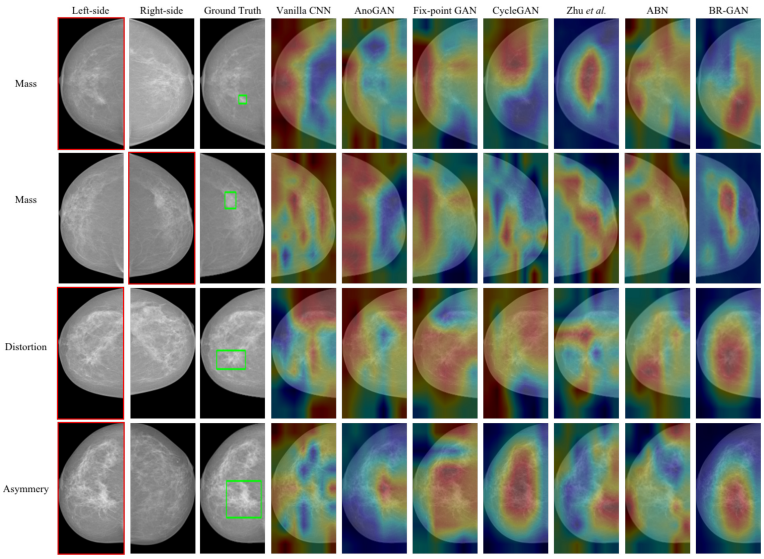

Fig. 3. Visualization of class activation maps of Vanilla CNN, AnoGAN [13], Fixed-Point GAN [14], CycleGAN [19], Zhu *et al.* [20], ABN [2] and our BR-GAN. The target containing lesions is bounded by a red rectangle. The ground truth bounding boxes are labeled by green rectangles in the third column. (Color figure online)

BR-GAN succeeds to focus on all lesions since it incorporates the bilateral asymmetry prior and modifies the cycle mechanism. The other methods show uneven results without considering bilateral information.

3.3 Ablation Experiments

To verify the effectiveness of each component, we evaluate some variant models and show results in Table 3. Here are some interpretation for the variants:

SBF: Simple Bilateral Features. The bilateral features are combined and fed into the fusion layer directly;

Single: Only use the consistency loss L_c^T;

Double: Use both consistency losses L_c^T and L_c^C;

Mask: Whether use attention mask in *Residual Preserved Module*.

Note that bilateral breasts exist misalignment, using SBF to classify is not robust enough. As shown in the above tables, the bilateral cycle mechanism, the double consistency losses, the residual preserved mechanism and attention mask for preserved features are all proved to be effective.

Table 3. Ablation experiments on (a) INBreast dataset for mass classification with AlexNet; (b) INBreast dataset for mass classification with ResNet50; (c) INBreast dataset for mixed-lesion classification with ResNet50; (d) in-house dataset for mixed-lesion classification with AlexNet.

Bilateral	L_C	L_{RE}	Mask	AUC (a)	AUC (b)	AUC (c)	AUC (d)
×	×	×	×	0.820	0.827	0.780	0.697
SBF	×	×	×	0.862	0.858	0.807	0.721
GAN	×	×	×	0.883	0.873	0.857	0.731
GAN	Single	×	✓	0.861	0.859	0.826	0.727
GAN	Double	×	✓	0.886	0.864	0.846	0.767
GAN	Double	✓	×	0.889	0.857	0.846	0.761
GAN	Double	✓	✓	**0.900**	**0.886**	**0.860**	**0.770**

4 Conclusions

In this paper, we present a novel approach called bilateral residual generating adversarial network (BR-GAN) to improve the mammogram classification performance. The approach proposes a novel way to generate the healthy version of target features to help find the abnormal features. Thus, BR-GAN enhances the interpretability of results for clinical diagnosis. Experimental results indicate that the proposed BR-GAN achieves the state-of-the-art in both the public and the in-house dataset.

Acknowledgement. This work was supported by MOST-2018AAA0102004, NSFC-61625201 and ZheJiang Province Key Research & Development Program (No. 2020C03073).

References

1. Dhungel, N., Carneiro, G., Bradley, A.P.: The automated learning of deep features for breast mass classification from mammograms. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 106–114. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_13
2. Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention branch network: learning of attention mechanism for visual explanation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714 (2019)
3. Haarburger, C., et al.: Multiparametric magnetic resonance image synthesis using generative adversarial networks. In: VCBM (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Hu, X., Jiang, Y., Fu, C.W., Heng, P.A.: Mask-ShadowGAN: learning to remove shadows from unpaired data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2472–2481 (2019)

6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
7. Li, H., Chen, D., Nailon, W.H., Davies, M.E., Laurenson, D.: A deep dual-path network for improved mammogram image processing. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1224–1228. IEEE (2019). https://doi.org/10.1109/ICASSP.2019.8682496
8. Lotter, W., Sorensen, G., Cox, D.: A multi-scale CNN and curriculum learning strategy for mammogram classification. In: Carneiro, G., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 169–177. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_20
9. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012). https://doi.org/10.1016/j.acra.2011.09.014
10. Nizan, O., Tal, A.: Breaking the cycle-colleagues are all you need. arXiv preprint arXiv:1911.10538 (2019)
11. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076
12. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. **8**(1), 4165 (2018). https://doi.org/10.1038/s41598-018-22437-z
13. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
14. Siddiquee, M.M.R., et al.: Learning fixed points in generative adversarial networks: from image-to-image translation to disease detection and localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 191–200 (2019)
15. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics. CA Cancer J. Clin. **69**(1), 7–34 (2019). https://doi.org/10.3322/caac.21551
16. Tai, S.C., Chen, Z.S., Tsai, W.T.: An automatic mass detection system in mammograms based on complex texture features. IEEE J. Biomed. Health Inf. **18**(2), 618–627 (2013). https://doi.org/10.1109/JBHI.2013.2279097
17. Wu, E., Wu, K., Cox, D., Lotter, W.: Conditional infilling GANs for data augmentation in mammogram classification. In: Stoyanov, D., et al. (eds.) RAMBO/BIA/TIA -2018. LNCS, vol. 11040, pp. 98–106. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00946-5_11
18. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
20. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 603–611. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_69