Bilateral Asymmetry Guided Counterfactual Generating Network for Mammogram Classification

Churan Wang, Jing Li, Fandong Zhang, Xinwei Sun, Hao Dong, *Member, IEEE*, Yizhou Yu, *Fellow, IEEE*, and Yizhou Wang

***Abstract*— Mammogram benign or malignant classification with only image-level labels is challenging due to the absence of lesion annotations. Motivated by the symmetric prior that the lesions on one side of breasts rarely appear in the corresponding areas on the other side, we explore to answer a *counterfactual question* to identify the lesion areas. This counterfactual question means: *given an image with lesions, how would the features have behaved if there were no lesions in the image?* To answer this question, we derive a new theoretical result based on the symmetric prior. Specifically, by building a causal model that entails such a prior for bilateral images, we identify to optimize the distances in distribution between i) the counterfactual features and the target side's features in lesion-free areas; and ii) the counterfactual features and the reference side's features in lesion areas. To realize these optimizations for better benign/malignant classification, we propose a counterfactual generative network, which is mainly composed of Generator Adversarial Network and a *prediction feedback mechanism*, they are optimized jointly and prompt each other. Specifically, the former can further improve the classi?cation performance by generating counterfactual features to calculate lesion areas. On the other hand, the latter helps counterfactual generation by the supervision of classification loss. The utility of our method and the effectiveness of each module in our model can be verified by state-of-the-art performance on INBreast and an in-house dataset and ablation studies.**

Manuscript received September 26, 2020; revised April 12, 2021 and July 26, 2021; accepted August 22, 2021. Date of publication September 17, 2021; date of current version September 23, 2021. This work was supported in part by the Grant MOST-2018AAA0102004, in part by the Grant NSFC-61625201, in part by Zhejiang Province Key Research and Development Program under Grant 2020C03073, and in part by Beijing Municipal Science and Technology Planning Project under Grant Z201100005620002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ajmal S. Mian. *(Chu-ran Wang and Jing Li contributed equally to this work.) (Corresponding authors: Xinwei Sun; Yizhou Wang.)*

Churan Wang and Fandong Zhang are with the Center for Data Science, Peking University, Beijing 100871, China (e-mail: churanwang@pku.edu.cn; fd.zhang@pku.edu.cn).

Jing Li is with the Department of Computer Science, Peking University, Beijing 100871, China (e-mail: lijingg@pku.edu.cn).

Xinwei Sun is with Peking University, Beijing 100871, China (e-mail: sxwxiaoxiaohehe@pku.edu.cn).

Hao Dong is with the Center on Frontiers of Computing Studies, Department of Computer Science, Peking University, Beijing 100871, China (e-mail: hao.dong@pku.edu.cn).

Yizhou Yu is with the Department of Computer Science, The University of Hong Kong, Hong Kong, and also with the AI Lab, Deepwise Healthcare, Beijing 100080, China (e-mail: yizhouy@acm.org).

Yizhou Wang is with the Department of Computer Science, Peking University, Beijing 100871, China, and also with the Center on Frontiers of Computing Studies, Peking University, Beijing 100871, China (e-mail: yizhou.wang@pku.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3112053

***Index Terms*— Domain Knowledge, Bilateral Asymmetry, Counterfactual, Mammogram Classification.**

I. Introduction

BREAST cancer is the leading cause of cancer death among women [35]. The mammography-based **B**enign/**M**alignant **C**lassification (BMC) is considered to be an effective way for early breast cancer diagnosis. Since the annotations of lesion areas are manually costly (*e.g.*, labelling bounding boxes [8], [23], [30], [36], [40] or binary masks [7]), it is desired for clinical use to achieve BMC with only image-level labels. The key towards this goal lies in the exploration of abnormal patterns/features, such as masses, calcification clusters, structure distortions and their associated signs like skin retraction, skin thickening and so on. However, the high-intensity breast tissues in 2D image (as projection of the 3D organ) may partially obscure the lesions, making the identification of the above abnormalities challenging.

To solve this problem, existing works mainly utilize specific rules or attention modules for feature selection. For example, [46] selects the local features with the maximum response or largest prediction score; [9], [44] selects the most discriminative region via the proposed attention branch supervised by a classification signal. However, these methods fail to consider the mammogram domain knowledge, which can be very valuable for lesion localization.

One important mammogram medical prior is "Anotomical Symmetry", which has been authenticated by BI-RADS standard of American College of Radiology [33]. Such a prior is two-fold. On one hand, the lesion area in the one side rarely appears in the corresponding area in the other, as illustrated by the unhealthy cases in Fig. 1 (a). On the other, for lesion-free areas in both sides, the bilateral breasts often share similar parenchymal texture, as shown in Healthy Breasts cases in Fig. 1 (b). This symmetric prior is also found in previous studies about breast cancer [2], [5], [21].

Inspired by such a prior, it is natural to ask the following counterfactual generation question: *what would the features of the target image have been looked like had lesions removed, given observed target image with lesions and the*

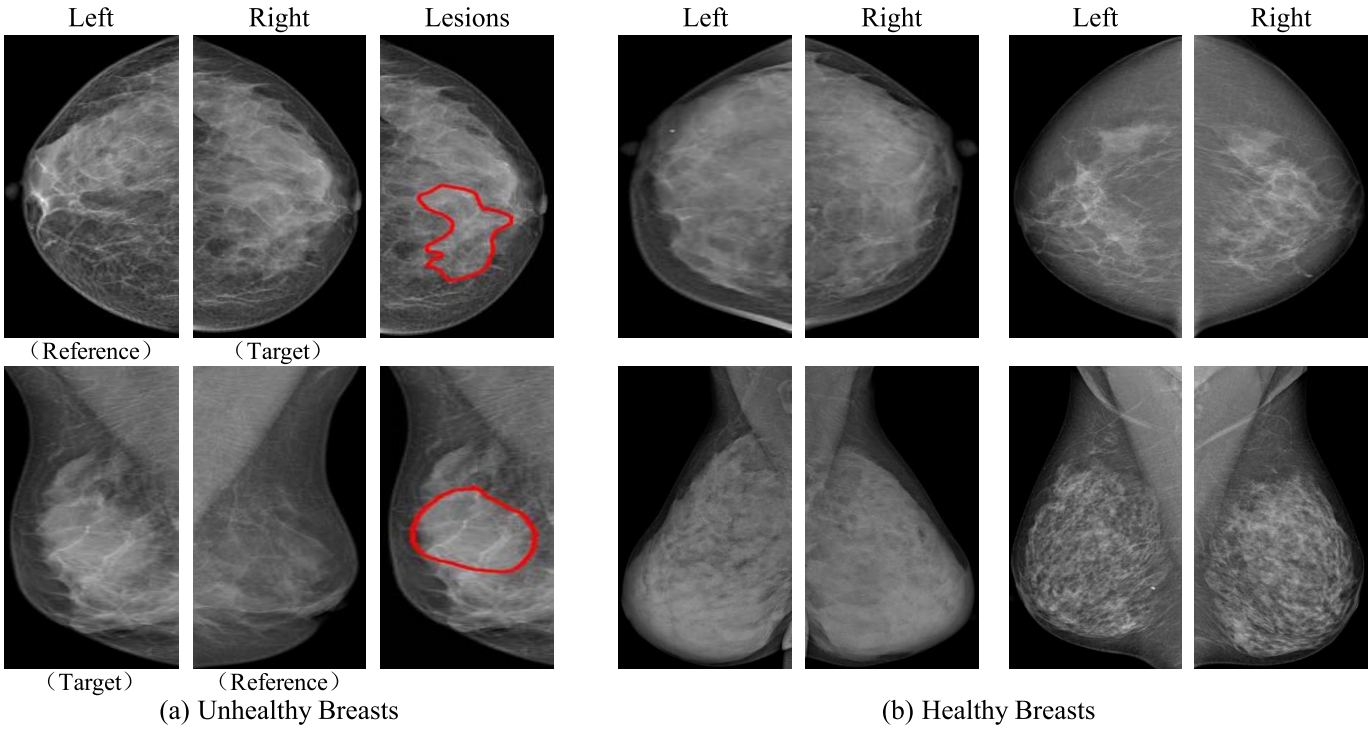


Fig. 1. (a) Two cases to show how the unhealthy breasts look asymmetrical. (b) Illustrations of that healthy breasts are roughly bilaterally symmetrical, with patterns and appearance (*e.g.,* structure, distribution, density, and morphology) of breast tissues can be very diverse among them.

reference image that is lesion-free in the corresponding area? After such counterfactual features being generated, the residue between the original target features and the counterfactual one incorporates the information of lesions. Hence the calculated residue can provide clinically explainable information for BMC. The answer to the above question is via constructing a structural causal model [28] in which the counterfactual learning is well defined. Specifically, a structural causal model (SCM) is proposed that introduces latent bilateral variables for generating bilateral images. To depict the bilateral symmetry, we further introduce a hidden confounder (including DNA, environment, etc.) that generates such bilateral features via the same causal mechanism. Following the symmetric prior, we can obtain an inspiring theory dubbed as *counterfactual constraints*: the target features of counterfactual generation share the same distribution (i) with the reference features in lesion areas and (ii) with the target features in lesion-free areas. Based on such a theoretical finding, we propose a novel **C**ounterfactual **G**eneration **N**etwork (CGN). Note that pixel-to-pixel registration between bilateral images is challenging due to unpleasant spatial distortion during image capturing and imperfect anatomical symmetry, we apply counterfactual generation in feature level motivated by [22]. Moreover, it achieves faster training speed without losing prediction power. This is also the reason why many domain adaptation methods work on feature space. Our CGN simultaneously optimizes the counterfactual generation (under counterfactual constraints) and lesion-area estimation via an attention-based prediction feedback mechanism, in an iterative manner. During optimization, both the lesion-area estimation and counterfactual generation prompt each other, supervised by classification loss.

In contrast to the existing GAN-based works [31], [34], [38], [45] for counterfactual generation, our method is endowed with a theoretical guarantee regarding the counterfactual distribution [6] by exploiting the symmetric prior. Specifically, AnoGAN [31] learns the latent space of healthy data and assumes that the lesions cannot be reconstructed within such latent space. Therefore the areas with large reconstruction errors are more likely to be lesions. Its performance highly depends on how well the healthy data is modeled. However, in our mammogram application, the glandular structure and characterization of healthy images can be very diverse. Sometimes the healthy pattern can even be similar to lesions, as shown in Fig. 1. Thus it is challenging to model healthy patterns well and distinguish the lesions at the same time using only healthy data. Although the cycle consistency loss [34], [45] can utilize the lesion information by learning a back translation (*i.e.,* from the counterfactual to the original), they also suffer from the healthy modeling problem in the forward translation (*i.e,* from the original to the counterfactual). What is more, these methods all assume that the translated data can be translated back to the original data [16], [26]. In our application, it means the back translation network should be able to model the location and appearance of the removed lesion. However, mammogram lesions can appear anywhere, *i.e.,* the location of the lesions is unpredictable. Therefore, it is an ill-posed problem to translate the counterfactual data back to the corresponding original data perfectly. BR-GAN [38] further improves cycle consistency mechanism by utilizing the information of the contralateral information during generation. However, this method implicitly assumes that the contralateral side contains no lesions, which may not hold in real scenarios.

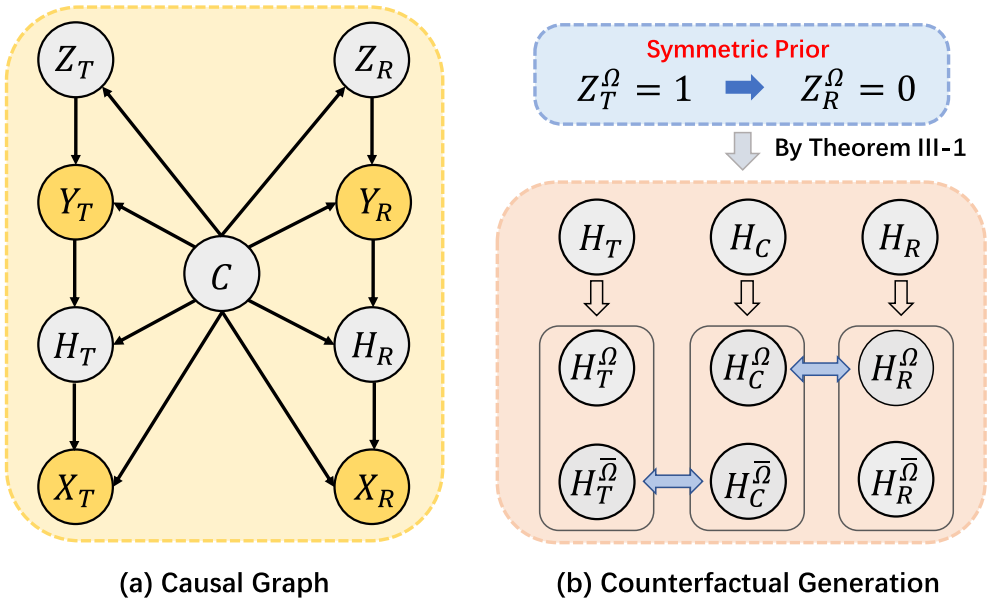


Fig. 2. (a): Our causal graph with observed variables marked by yellow and unobserved variables marked by gray. For notations, C denotes the DNA, growth environment that can explain the common properties shared between X_T and X_R; Z_R, Z_T denote lesion states ($Z^{\Omega}_{u=T,R} = 1$ if there are lesions in Ω; and $= 0$ if not); H_R, H_T respectively denote the hidden features of the image. (a) is mathematically expressed in our Eq (1). (b): Our counterfactual learning framework, motivated by symmetric prior (as shown in the top blue box). Our theoretical result (theorem 1) is illustrated in the bottom orange box, in which the H_C denotes the counterfactual result of the target side with the removal of lesion areas, *i.e.,* the counterfactual result of H_T under counterfactual event $Z_T = 0$. The blue arrows denote "distributionally equivalence". As shown, the distribution of H_C^{Ω} is the same with H_R^{Ω}, described by Eq. (2); the distribution of $H_C^{\overline{\Omega}}$ is the same with $H_T^{\overline{\Omega}}$, described by Eq. (3).

Instead of learning from healthy images, our CGN applies counterfactual generation conditioning on the bilateral information. Based on the symmetry prior, we propose to generate counterfactual features and estimate lesion areas together under counterfactual constraints: being similar in distribution with the reference features in lesion areas and maintaining the information of the target features in lesion-free areas. The whole pipeline of our method is illustrated in Fig. 3. Specifically, we first apply a deep generator with AdaIN [18] mechanism to provide the feature generation ability. Then we design a prediction feedback mechanism to help estimate the lesion areas. Meanwhile, an adversarial reference loss, a feedback triplet loss, and an auxiliary negative embedding loss are proposed to encourage the generated features to satisfy the above counterfactual constraints. Both the lesion-area estimation and counterfactual generation are optimized jointly and prompt each other. Further, we get the residual features by computing the difference between the generated counterfactual features and target features. Finally, we aggregate the residual features together with the target features for the final classification.

We evaluate the proposed method on a public dataset INBreast [25] and an in-house dataset. Our CGN achieves an area under the curve (AUC) of 91.3% on INBreast and 78.1% on the in-house dataset, which largely outperforms the representative methods. Our contributions are summarized as follows:

1) **Ideologically,** we exploit the symmetric prior into counterfactual generation for benign/malignant classification.
2) **Theoretically,** we prove that the counterfactual features should follow the *counterfactual constraints*, under the symmetric prior.
3) **Methodologically,** we propose the counterfactual generation network, guided by the counterfactual constraints.
4) **Experimentally,** we achieve state-of-the-art performance for mammogram classification on both the public and the in-house datasets.

II. Related Work

A. Counterfactual Generation

Existing GAN-based models for counterfactual generation can be roughly categorized into two classes: (i) healthy modeling methods, *e.g.,* AnoGAN [31] and (ii) methods based on cycle consistency, *e.g.,* CycleGAN [45], Fixed-point GAN [34]. For class (i), they propose to model the pattern of healthy data. However, these methods suffer from unstable result due to large diversity of glandular structure and characterization of healthy images. For class (ii), they use cycle consistency loss to incorporate bi-directed translation: forward translation (from the original to the counterfactual) and back translation (from the counterfactual to the original). These methods suffer from two problems: a) the healthy modeling problem for forward translation, similar to class (i); b) the ill-posed problem for back translation since the location and appearance of the removed lesion are diverse and unpredictable. In comparison with existing works, our method learns healthy pattern by exploiting symmetric prior, so as to avoid the problems mentioned above. As a result, our model is able to achieve more robust counterfactual generation result.

B. BMC With Only Image-Level Labels

Previous approaches that can be used to address BMC with only image-level labels without any extra annotations

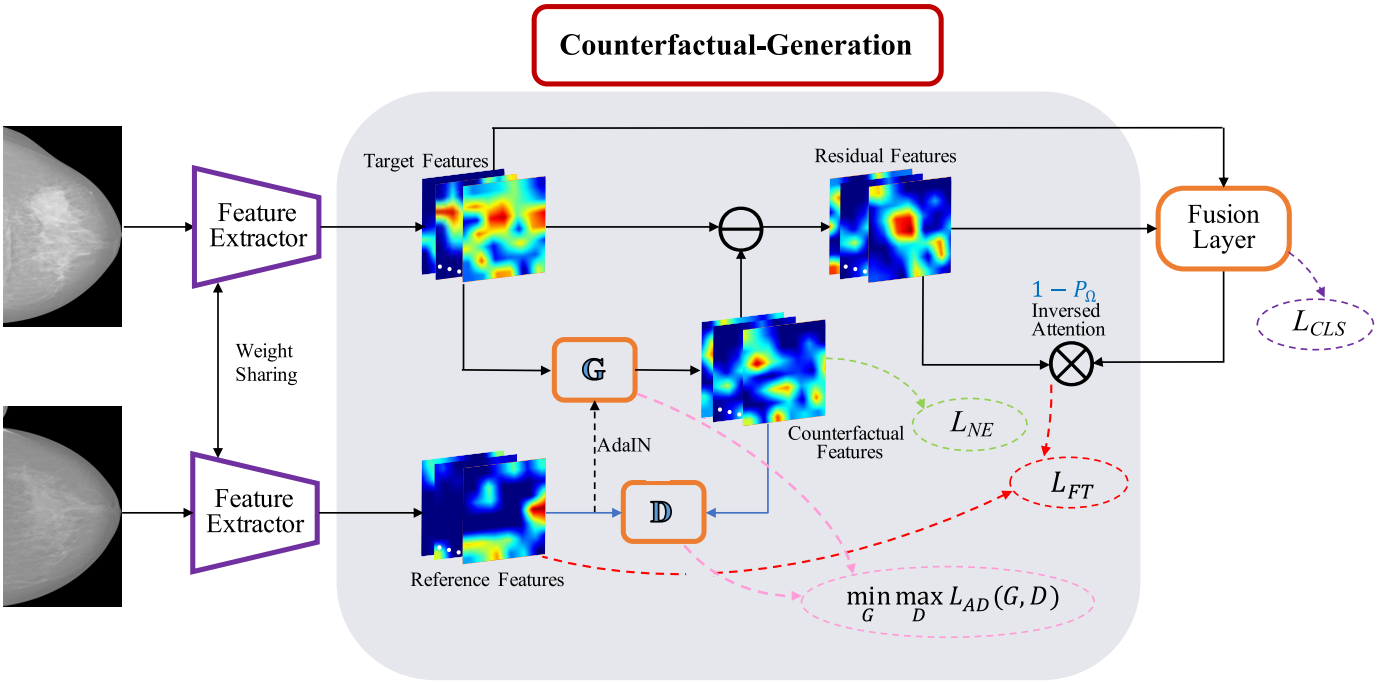


Fig. 3. The schematic overview of CGN. First, two feature extractors with weight sharing extract the features for input paired target and reference images, respectively. Then the bilateral features are processed by AdaIN mechanism and fed into the generator G to generate the counterfactual features. The counterfactual features are constrained by adversarial learning with a feedback triplet loss $\mathcal{L}_{FT}$, and a negative embedding loss $\mathcal{L}_{NE}$. Then, the residual features are obtained by computing the difference between the target features and counterfactual features. Finally, the residual features are fed into a *Fusion* network with target features and outputs prediction of benign/malignant.

are roughly categorized into two classes: (i) the attention-based methods, *e.g.,* Zhu *et al.* [46], Zhou *et al.* [44] and Fukui *et al.* [9]; (ii) the multi-view fusion methods, *e.g.,* Wu *et al.* [41], BR-GAN [38]. For class (i), they append a response-based visual explanation model with an attention module or specific rules. However, they all ignore medical domain knowledge which is valuable for BMC and are fragile when facing dense breasts without learning from bilateral information. For class (ii), since the bilateral breasts are not pixel-to-pixel symmetry, direct multi-view fusions [41] can be very sensitive to bilateral misalignment. Although the residual-preserved mechanism proposed by [38] could improve the cycle consistency mechanism, the location of lesions can be inaccurate due to lack of leveraging symmetric prior regarding lesion areas. In contrast, our method can avoid the problems above by taking advantage of the symmetric prior elaborately. More importantly, our method is guaranteed by theoretical formulation on the counterfactual generation via structural causal model.

III. Methodology

Problem Setup and Notations: The goal of mammogram benign or malignant classification is to learn classifier $f : \mathcal{X} \to \mathcal{Y}_T$ that predicts the disease label of target side X_T. Where $\mathcal{X} := (\mathcal{X}_T, \mathcal{X}_R)$ $(\mathcal{X}_T, \mathcal{X}_R \subset \mathbb{R}^d)$ denotes the input space of bilateral breast images. Here, T and R respectively stand for the target and reference side of bilateral breast images. $\mathcal{Y}_T := \{0, 1\}$ denotes the disease label of the target side (1 denotes malignant and 0 denotes benign). To achieve this goal, we are given training data $\{(x_T^i, x_R^i, y_T^i)\}_{i \in [N]}$ $([N] := \{1, \ldots, N\}$ for any integer $N > 0)$. During test stage, our goal is to predict y_t for a new instance $x = (x_T, x_R) \in \mathcal{X}$.

A. Counterfactual Learning

Symmetric Prior [33]: For a paired image data, if the target image contains lesions, the corresponding symmetrical area in the reference image has almost certainly no lesions. If the area on both sides contain no lesions, they share the same distribution on parenchymal texture.

This symmetric prior provides guidance for localizing lesion areas, as a residue between the features of the target image and the one with the removal of corresponding lesions. The generation of the latter image is a counterfactual question, *i.e., what would the features of target image have been looked like had lesions removed, given observed target image with lesions and the reference image that is lesion-free in the corresponding area?* Such a counterfactual question has been well-defined and explored in the framework of Structural Causal Model (SCM) [28] that entails the causal assumptions in the corresponding causal graph. In our scenario, these causal assumptions come from the symmetric prior between bilateral images.

To describe this bilateral symmetry, we propose a SCM that introduces a hidden common factor (denoted as C which can refer to DNA, growth environment, etc.) as the confounder of the bilateral variables. This bilateral generation depicts our symmetric prior as shown in Fig. 2 (a). Besides, our SCM incorporates bilateral latent features $H_{U=T,R}$, as abstraction/concepts of bilateral images. Such bilateral features, which are affected by C and disease status $(Y_{U=T,R})$

that is determined by lesion status $Z_{U=T,R}$. The distribution of these variables are assigned by the following structural equations:

$$C = f_C(\epsilon) \rightarrow \begin{cases} Z_T = f_{Z_T}(C) \\ Z_R = f_{Z_R}(C) \end{cases} \rightarrow \begin{cases} Y_T = f_{Y_T}(C, Z_T) \\ Y_R = f_{Y_R}(C, Z_R) \end{cases}$$
$$\rightarrow \begin{cases} H_T = f_{H_T}(C, Y_T) \\ H_R = f_{H_R}(C, Y_R) \end{cases} \rightarrow \begin{cases} X_T = f_{X_T}(C, H_T) \\ X_R = f_{X_R}(C, H_R). \end{cases} \quad (1)$$

Equipped with such a SCM, we can mathematically formulate the symmetric prior as $Z_T^{\Omega} = 1 \rightarrow Z_R^{\Omega} = 0$, with Ω denoting the lesion areas of the target image X_T. The counterfactual generation question can be formulated as ${H_{T(Z_T=0)}}^{\Omega}(c)$ that can be read as the value of H_T on Ω in situation $C = c$ had $Z_T^{\Omega} = 0$ [28]. Since the situation $C = c$ is induced by the factual event $\{H_T^{\Omega} = h_t, Z_T^{\Omega} = 1\}$, our counterfactual distribution can be denoted as $P({H_{T(Z_T^{\Omega}=0)}}^{\Omega} = h | H_T^{\Omega} = h_t, Z_T^{\Omega} = 1)$. Under our SCM and the symmetric prior, we have following results for counterfactual generation:

Theorem 1: Under the symmetric prior, the structural equation model defined in Eq. (1) *for Fig. 2 (a) has the following results for counterfactual distribution of target features:*

$$P({H_{T(Z_T^{\Omega}=0)}}^{\Omega} = h | H_T^{\Omega} = h_t, Z_T^{\Omega} = 1)$$
$$= P(H_R^{\Omega} = h_r | H_T^{\Omega} = h_t, Z_T^{\Omega} = 1) \quad (2)$$
$$P({H_{T(Z_T^{\overline{\Omega}}=0)}}^{\overline{\Omega}} = h | H_T^{\overline{\Omega}} = h_t, Z_T^{\overline{\Omega}} = 0)$$
$$= P(H_T^{\overline{\Omega}} = h_t | H_T^{\overline{\Omega}} = h_t, Z_T^{\overline{\Omega}} = 0), \quad (3)$$

The proof of Theorem 1 is shown in our appendix. This theorem implies the following *counterfactual constraints*: the distribution of generated counterfactual features should be equal (i) to reference features in lesion areas, (ii) to target features in lesion-free areas. To realize the above counterfactual constraints, we propose to optimize the following objectives for the counterfactual generation:

$$\min_{\theta} \quad D(P_\theta({H_{T(Z_T^{\Omega}=0)}}^{\Omega}), P_\theta(H_R^{\Omega})) \quad (4)$$

$$\min_{\theta} \quad D(P_\theta({H_{T(Z_T^{\overline{\Omega}}=0)}}^{\overline{\Omega}}), P_\theta(H_T^{\overline{\Omega}})), \quad (5)$$

where D denotes generalized distance measure, *e.g.,* KL divergence. With such counterfactual learning, it is expected that the lesion areas, as the subtraction of counterfactual generation of H_T (with lesions removed) from original H_T, can be detected precisely and hence can lead to accurate classification performance. To achieve the above two goals, we propose a counterfactual generating network (CGN), which cooperatively localizes the lesion areas and achieve counterfactual generation simultaneously. We explain the CGN in details in the subsequent section.

B. Counterfactual Generating Network (CGN)

As illustrated in Fig. 3, our counterfactual generation network for mammogram classification contains the following steps: (i) Extract the target and reference features H_T and H_R from images X_T and X_R, via a feature extractor chosen from backbone network, *e.g.* AlexNet [19], ResNet [14], DenseNet [17], EfficientNet [37], (ii) a ***counterfactual generation module*** is designed to generate counterfactual features H_C from both H_T and H_R, (iii) ***a classification module*** is designed to predict malignant/benign, with aggregated H_C and H_T as input. To accurately identify Ω for generating H_C in step (ii), a ***prediction feedback mechanism*** and a set of ***counterfactual constrains*** motivated by Eq. (4) and (5) are designed. In what follows, we will explain the above mechanisms in more details.

1) Counterfactual Generation Module: The Adaptive Instance Normalization (AdaIN) [18] has been proved to be effective for style transfer tasks. It is adopted in the generator G (as shown in Fig. 3) for counterfactual generation, with H_T as content and H_R as style in our case:

$$AdaIN(H_T, H_R) = \sigma(H_R)\left(\frac{H_T - \mu(H_T)}{\sigma(H_T)}\right) + \mu(H_R), \quad (6)$$

with $\mu(\cdot)$ and $\sigma(\cdot)$ denoting the mean and standard variance function. As suggested by [18], an interpolated H_T and AdaIN are fed into a generator network containing nine residual blocks to generate counterfactual features H_C:

$$H_C = G((1-\alpha) * H_T + \alpha * AdaIN(H_T, H_R)), \quad (7)$$

where α is a hyper-parameter of the interpolation weight.

2) Classification Module: The residual features (entailing lesion information) obtained by $H_T - H_C$ and H_T (with additional contextual information which is showed useful for the medical image inference [3] besides lesion-related information we obtained) are fed into a classifier in a concatenated way. This classifier, which implements a convolutional block as FusionLayer to obtain the fused features, is trained via commonly used cross-entropy loss:

$$\mathcal{L}_{CLS}(G) = -(Y_T^{gt} * \log Y_T(H_T - G(H_R, H_T), H_T) + (1 - Y_T^{gt}) * \log(1 - Y_T(H_T - G(H_R, H_T), H_T))), \quad (8)$$

where $Y_T(H_T - G(H_R, H_T), H_T)$ (with $H_C = G(H_R, H_T)$) is the classification probability of being malignant.

3) Prediction Feedback Mechanism: This mechanism is to estimate the lesion areas Ω for follow-up counterfactual generation (specifically Eq. (10)). For implementation, we use the attention mechanism, in which the attention map is calculated by normalization/softmax following the class activation map (CAM) [44], *i.e.,* $P_\Omega = softmax(CAM)$ that denotes the probability map of being lesions. Higher value of a pixel implies higher probabilities of being lesioned. Thus $1 - P_\Omega$ measures the probability of not being lesioned.

4) Counterfactual Constraints: Since the direct optimization of Eq. (4) and (5) can be intractable/unstable for general distance measure D such as KL-divergence, we adopt the adversarial learning strategy [10]. For optimization of Eq. (4), GAN generates similar features from the whole reference image and can constrain our desired features be the same as the reference in lesion areas. Specifically, a Discriminator D (learns to classify H_C and H_R) and a Generator G (fools the discriminator) are designed and trained in a competing way:

$$\min_G \max_D \mathcal{L}_{AD}(G, D) := \log(D(H_R)) + \log(1 - D(G(H_T, H_R))). \quad (9)$$

However, the generated features through GAN loss are undesired features in lesion-free areas. For optimization of Eq. (5), we use a prediction feedback mechanism to localize lesion areas. One intuitive way to use feedback mechanism is constraining generated features to be the same as the target/reference features in lesion-free/lesion areas directly. However, motivated by [32], the triplet loss can be better than such designs. They will suffer from slow convergence and falling into local minimum easily. We analysis and evaluate such variant methods in Sec IV-F. Thus, we propose a feedback triplet loss to minimize the distance between the target features $H_T^{\overline{\Omega}}$ and counterfactual features $H_C^{\overline{\Omega}}$ in lesion-free areas, which is measured by target-counterfactual distance d_{tc} by weighted mean square error:

$$d_{tc} = \frac{\sum_i^h \sum_j^w (1 - P_\Omega^{ij}) \left\| H_T^{ij} - H_C^{ij} \right\|_2^2}{h \times w - 1}, \tag{10}$$

where h and w denote the height and width of CAM respectively. Motivated by minimization of distance between H_C and H_R enforced by Eq. (9), we choose a d_{rc} between H_R and H_T as an adaptive reference to minimize d_{tc}. The d_{rc} is measured by chamfer distance [1] to endure the misalignment, and is defined by

$$d_{rc} = \frac{\sum_i^h \sum_j^w (\underset{u,v}{min} \left\| H_R^{ij} - H_C^{uv} \right\|_2^2 + \underset{u,v}{min} \left\| H_C^{ij} - H_R^{uv} \right\|_2^2)}{2 \times h \times w}. \tag{11}$$

Therefore, the feedback triplet loss is defined as:

$$\mathcal{L}_{FT}(G) = \max\{0, d_{tc} + \beta - d_{rc}\}. \tag{12}$$

The triplet loss makes H_C be closer to H_T than H_R in terms of the lesion-free areas. Further the GAN loss makes the distance between H_C and H_R be close in the lesion areas. Based on the cooperation of GAN loss and the triplet loss, the generated H_C satisfies Eq. (4) and (5). Besides, $L_{FT}(G)$ as a margin term can avoid learning identity mapping from H_T to H_C during minimizing $L_{FT}(G)$. Catering misalignment is not needed for d_{tc} since H_C is for the "target" and hence perfectly aligned with H_T in pixel-wise.

Besides, since the lesion regions of H_T have been removed in H_C, the H_C must also be non-malignant. Such a knowledge can be reflected via an auxiliary negative embedding loss as a constraint:

$$\mathcal{L}_{NE}(G) = -\log(1 - p_m(H_C)), \tag{13}$$

where $p_m(H_C)$ denotes the malignant probability of H_C.

5) Joint Optimization: The final loss is combination of the losses defined in Eq. (8), (9), (12) and (13):

$$\min_G \max_D \mathcal{L}(G, D) := \sum_k \left\{ \mathcal{L}_{AD}^k(G, D) + \mathcal{L}_{NE}^k(G) + \mathcal{L}_{FT}^k(G) + \mathcal{L}_{CLS}^k(G) \right\}, \tag{14}$$

where k denotes sample index, that is, we calculate corresponding losses for each sample and derive the final joint loss. By optimizing the loss $\mathcal{L}(G, D)$, these modules can be optimized cooperatively and compatibly: the counterfactual generation helps discover the lesions for classification; on the other hand, the classification module helps counterfactual generation in a supervised way. The effect of these modules can be validated by our ablation study, which are explained detailedly in the next section.

IV. Experiments

A. Datasets

We evaluate our method on both the public and the in-house datasets, with only image-level malignant/benign labels are provided. For benign/malignant classification, we select only lesional images according to the clinical report, as lesion-free images are healthy by default. For the public dataset, we use INBreast dataset [25] due to its high quality compared to other public datasets [46] and an in-house dataset. The INBreast dataset contains 115 cases and 410 mammograms. INBreast provides each image a BI-RADS result as image-wise ground truth and we use the same pre-prosessing method of labels as Zhu *et al.* [46](malignant if BI-RADS > 3; benign otherwise). We share the same dataset with Zhu *et al.* [46] in which the dataset contains 100 mammogram images with masses, except that we discard 9 of them for lack of contralateral images. We use five-fold cross-validation for evaluation and area under the curve (AUC) for measurement.

The in-house dataset contains 2500 images. We select 1303 images of them that contain image-level benign or malignant labeling annotations. The selected dataset contains 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 (Note that the target image A with a malignancy annotation is paired with B, counting as one pair. Meanwhile, if B also has a malignancy annotation, conversely B can be the target and A can be the reference, counting as another one pair). We randomly divide the dataset into training, validation and testing sets by the proportion of $8:1:1$ in patient-wise.

B. Experiment Settings

To validate the utility of our method, we apply our model on two problem settings: the mass-lesion image classification followed by [46] and mixed-lesion classification in which the lesions can be masses, calcification clusters and distortions. We adopt various backbones in each setting. All together, we implement our model on: mass malignancy classification on INBreast dataset with AlexNet [19], Resnet50 [13], DenseNet [17] and EfficientNet [37] as backbone respectively; mixed-lesion malignancy classification with Resnet50 [13] as backbone on INBreast dataset; and mixed-lesion malignancy classification with AlexNet [19] as backbone on in-house dataset. To make a fair comparison, all methods (our method and compared baselines) share the same backbone in each setting.

C. Implementation Details

Mammogram images are commonly stored using a 14-bit DICOM format. A simple linear mapping is used to convert

TABLE I

AUC EVALUATION OF COMPARATIVE EXPERIMENTS ON (a) INBREAST + ALEXNET (MASS); (b) INBREAST + RESNET50 (MASS); (c) INBREAST + DENSENET (MASS); (d) INBREAST + EFFICIENTNET (MASS); (e) INBREAST + RESNET50 (MIXED LESIONS); (f) IN-HOUSE + ALEXNET (MIXED LESIONS); NOTE THAT THE '*' MEANS OUR RE-IMPLEMENTATION. THE '-' MEANS THERE ARE NO OFFICIAL REPORT RESULTS

Experiment Setting	**AUC (Mass)**				**AUC (Mixed lesions)**	
Datasets	INBreast [25]				INBreast [25]	In-House data
Backbones	AlexNet [19]	ResNet [13]	DenseNet [17]	EfficientNet [37]	ResNet [13]	AlexNet [19]
Pretrained CNN [8]	0.690	–	–	–	–	–
Pretrained CNN+Random Forest [8]	0.760	–	–	–	–	–
Vanilla AlexNet in Zhu *et al.* [46]	0.790	–	–	–	–	–
Zhu *et al.* [46]	0.890	–	–	–	–	–
Vanilla*	0.820	0.827	0.822	0.830	0.780	0.697
AnoGAN [31]*	0.803	0.796	0.794	0.804	0.774	0.720
Fixed-Point GAN [34]*	0.835	0.837	0.831	0.841	0.805	0.734
CycleGAN [45]*	0.852	0.838	0.837	0.852	0.808	0.741
Wu *et al.* [41]	0.863	0.860	0.854	0.862	0.810	0.723
Zhu *et al.* [46]*	0.860	0.862	0.852	0.863	0.830	0.720
Vanilla*+GAP [44]*	0.857	0.827	0.822	0.830	0.780	0.718
Vanilla*+ABN [9]*	0.858	0.846	0.833	0.849	0.814	0.723
BR-GAN* [38]	0.900	0.886	0.884	0.887	0.860	0.770
Proposed Method	**0.910**	**0.911**	**0.908**	**0.913**	**0.885**	**0.781**

them into 8-bit gray images. Then, the Otsus method [27] is used for breast region segmentation and background removal. The segmented images are resized into 224×224 and fed to networks. For each training epoch, we follow the [46] to randomly flip the mammograms horizontally and rotate within 45 degrees. To maintain the symmetric prior, we concatenate the target and the reference image in channel-wise and implement random data augmentation in the same time. After data augmentation, we split the target image and the reference image in channel-wise. The models are initialized by ImageNet pre-trained weights, since the ImageNet model can extract high-level features of medical images, as used in [43]. Besides, as a large-scale dataset, the ImageNet pre-training has been validated to help optimization on small dataset and widely adopted for better training in medical imaging [11], [20], [29], [42], [46]. Such an implementation of ImageNet pre-training is also for a fair comparison with the baseline method [46] in Table I, which pretrains the model on ImageNet to improve the training efficiency on small-scale medical data.

We adopt the Adam optimization with a learning rate of $5e-5$. For each method, we train for 50 epochs and select the best model on the validation set for testing. Both target and reference features are extracted from the last convolution layer. We implement all models with PyTorch.

D. Compared Baselines

We conduct our experiments on both Mass malignancy classification (the 2nd to 5th columns of Table I) and Mixed-lesion malignancy classification (the last two columns of Table I). The compared baselines are: 1) **Pretrained CNN** [8]: pre-trains the CNN with the regression to the hand-crafted features; 2) **Pretrained CNN+Random Forest** [8]: uses the random forest as the classifier, with the last layer of the CNN pre-trained by 1) as input; 3) **Zhu *et al.*** [46] selects local features with the maximum response and diagnosis based on the largest prediction score; 4) **Vanilla**: trains the network (*i.e.,* AlexNet [19], ResNet50 [14], DenseNet [17] or EfficientNet [37]) via vanilla empirical risk minimization; 5) **AnoGAN** [31]: uses GAN to learn the latent space of healthy images, then the residue between which and that of the original image is used for classification; 6) **Fixed-Point GAN** [34]: generates healthy version of the target image via fixed-point translation learning, followed by classification on the residue between this generated healthy image and the original one; 7) **CycleGAN** [45]: replaces the fixed-point translation learning with the cycle consistency loss during generating the healthy image; 8) **Wu *et al.*** [41]: designs a deep simple four-view CNN for classification; 9) **Vanilla+GAP** [44]: incorporates Global Average Pooling (GAP) into Vanilla CNN in 4); 10) **Vanilla+ABN** [9]: incorporates Attention Branch Network (ABN) into Vanilla CNN in 4); 11) **BR-GAN** [38]: generates healthy features by referring the contralateral image based on cycle consistency and residual-preserved mechanism. The residue between the healthy features and original one is used for classification.

The 4th to 7th rows in Table I correspond to the official results reported in each representative methods. Due to the slightly difference in the number of images used by reference absence, for a fair comparison, we re-implement some baselines in the list such as vanilla methods (AlexNet [19]/ ResNet50 [14]/DenseNet [17]/EfficientNet [37]), mammogram classification methods [38], [41], [46], natural image classification methods [9], [44] and counterfactual generation methods [31], [34], [45].

E. Experimental Analysis

1) Result Analysis: As shown in Table I, our method performs better than others in all settings. Specifically, we outperform attention-based methods (Zhu *et al.* [46], ABN [9] and CAM [44]) largely by 4.9% to 10.5%, multi-view method (Wu *et al.* [41]) largely by 4.7% to 7.5% and GAN-based methods (AnoGAN [31], Fixed-Point GAN [34], CycleGAN [45] and Wang [38]) by 1.0% to 11.5%. Due to the attention mechanism, Zhu [46], ABN [9] and CAM [44] can

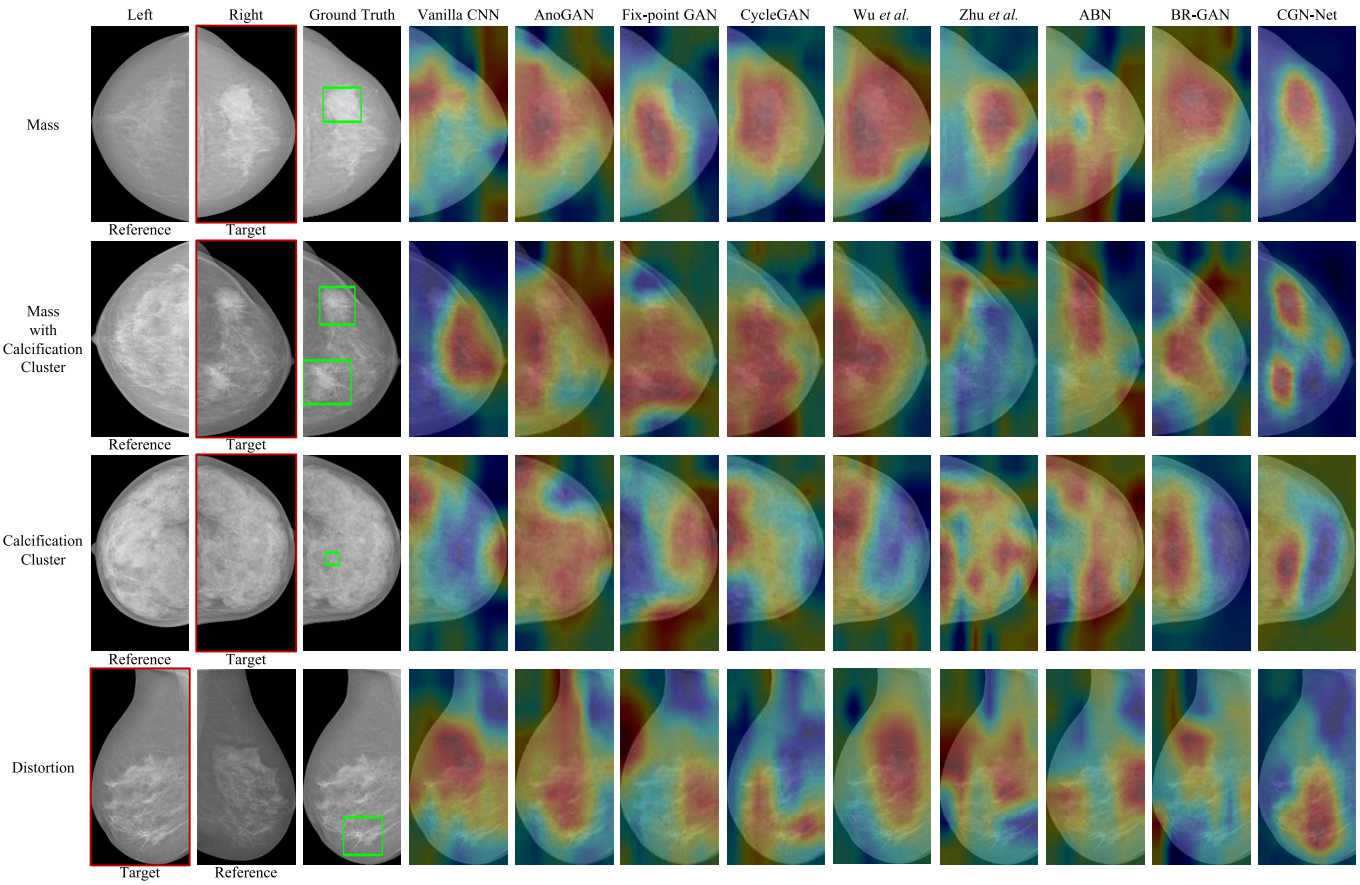


Fig. 4. Visualization of class activation maps of Vanilla CNN, AnoGAN [31], Fixed-Point GAN [34], CycleGAN [45], Wu *et al.* [41], Zhu *et al.* [46], ABN [9], BR-GAN [38] and CGN. Each row represents a pair of mammograms from bilateral breasts in the INBreast. The target containing lesions is bounded by a red rectangle. The ground truth bounding boxes are labeled by green rectangles.

outperform the vanilla baseline. However, without exploiting the domain knowledge of mammograms, their performances are limited. The improvement of Wu *et al.* [41] over the vanilla baseline indicates the benefit of leveraging bilateral information. Further, our method does not require pixel-to-pixel alignment that is inappropriate for mammogram images however was adopted by Wu *et al.* [41]. As to AnoGAN [31], compared with the vanilla baseline, AnoGAN performs slightly worse on INBreast dataset than on the in-house dataset. This may be due to the fact that the in-house dataset contains more healthy images. However, due to the difficulty of modeling a large variety of healthy patterns, these methods are still limited in terms of prediction power. Fixed-Point GAN [34] and CycleGAN [45] share the similar cycle consistency constraint and perform comparably. They outperform AnoGAN since they can make use of the image-level annotations. However, their performances are limited due to suffering from the ill-posed translation on lesion removal. Although BR-GAN [38] could alleviate this problem by incorporating the contralateral information during generation, it could not generalize to the case when the reference side contains lesions, leading to descent in performance compared with ours.

2) Localization Evaluation: To evaluate the effectiveness of our method in localizing lesion areas, we measure the localization error by CAM [44]. Similar to [44], we first calculate the CAMs based on the predicted category. Then we segment the regions whose CAM value is greater than 20% of the maximum CAM value and obtain the bounding box for the largest connected component in the segmentation map. We use the top-1 localization error adopted in ILSVRC, with the intersection over union (IOU) threshold set to 0.1.[1] As is shown in Table II, our CGN obtains a localization error of 0.421 for masses and 0.455 for all lesions, significantly outperforming other methods. In particular, our CGN can outperform BR-GAN by 8.9%-9.8% due to the appropriate modeling of symmetric prior.

3) Visualization: To verify the effectiveness of CGN in terms of learning lesion area, we visualize the class activation maps, as shown in Fig. 4. First, we can observe (the first three columns) the asymmetry in lesions' positions between bilateral images, which validates the bilateral asymmetric prior. Besides, our CGN succeeds to localize all lesions due to the incorporation of the bilateral symmetry prior. In contrast, the other methods can detect out-of-lesion regions, especially in the last two cases due to unobvious indistinct contrast between the lesions and tissues. Specifically, the redundant

[1]The original IOU threshold is 0.5, which is normally adopted in detection task. As our mainly focus on classification, we only require the localization to be roughly contain the lesion areas.

TABLE II

TOP-1 LOCALIZATION ERROR ON INBREAST DATASET FOR MASS CLASSIFICATION WITH RESNET50; INBREAST DATASET FOR MIXED-LESION CLASSIFICATION WITH RESNET50

Methodology	**Top-1 error** (Mass)	**Top-1 error** (Mixed lesions)
ResNet50[13]	0.635	0.727
AnoGAN [31]*	0.684	0.789
Fixed-Point GAN [34]*	0.646	0.737
CycleGAN [45]*	0.632	0.667
Wu *et al.* [41]*	0.627	0.650
ABN [9]	0.632	0.722
Zhu *et al.* [46]*	0.627	0.625
BR-GAN [38]	0.519	0.544
Proposed Method	**0.421**	**0.455**

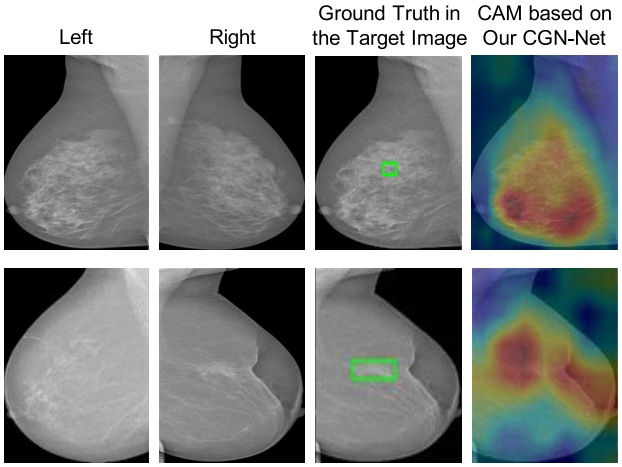


Fig. 5. Visualization of failure cases. The lesion in the first case is small amorphous calcification; the lesion in the second case is mass with associated signs of skin retraction.

areas selected by *BR-GAN* in the last case in Fig. 4 may be accounted by the lesions existed in the reference side.

4) Failure Case Visualization: Some failure cases are presented in Fig. 5. As shown in the first row, the lesions of which patterns are not obvious, such as microcalcifications; hence they may be overwhelmed by other asymmetrical regions during learning. As shown in the second row, under the symmetric prior, our method also learns additional asymmetrical areas which are highly suspicious lesion areas (*e.g.,* skin retraction). According to the American College of Radiology [33], the additionally learned retraction in the second case can be viewed as the associated signs caused by lesions, hence can also be beneficial for prediction.

F. Ablation Study

We evaluate some variant models to verify the effectiveness of each component. The ablative results in Table. III show that deleting or changing any of the components would lead to a descent of the classification performance. Specifically, naive bilateral features fusion also leads to a boosting of 2.4% to 4.2% over vanilla on performance. It proves that the bilateral symmetric prior is quite helpful for malignancy classification. Meanwhile, the proposed prediction feedback mechanism outperforms the non-feedback largely by 4.8%. We explain that the classification module provides additional useful supervision for lesion localization, making learning more accurate and stable. For additional counterfactual constraint of negative embedding loss, we show that it improves the performance by 1.1%. Some variants in our ablation studies are listed below:

$\times$ **in the first raw**: Vanilla single view network.

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

TF-GAN: Target-Feature GAN. Replace AdaIN input by target features only;

BF-GAN: Bilateral-Feature GAN. Replace AdaIN input by simple combination of bilateral features;

Non-feedback: Estimate lesion areas Ω by the areas with the largest target-counterfactual distance.

To further verify the effectiveness of the proposed adversarial loss and feedback triplet loss $\mathcal{L}_{FT}$, we implement two variants respectively:

Variant (1): As to the discriminator loss, we directly minimize the distance between counterfactual features H_C^{Ω} and reference features H_R^{Ω} in lesion areas. We still estimate the lesion areas Ω by the prediction feedback mechanism.

Compared with the competing losses we used for discriminator and generator in our method:

$$\min_G \max_D \mathcal{L}_{AD}(G, D) := \log\left(D\left(H_R\right)\right) + \log\left(1 - D\left(G(H_T, H_R)\right)\right), \quad (15)$$

we denote the modified discriminator loss and generator loss of variant (1) as:

$$\mathcal{L}_G^{\Omega} = log\left(1 - D\left(H_C^{\Omega}\right)\right) \quad (16)$$

$$\mathcal{L}_D^{\Omega} = -log\left(1 - D\left(H_C^{\Omega}\right)\right) - log\left(D\left(H_R^{\Omega}\right)\right) \quad (17)$$

therefore we have the final losses:

$$\mathcal{L}_1 = \mathcal{L}_G^{\Omega} + \mathcal{L}_{NE} + \mathcal{L}_{CLS} \quad (18)$$

which are iteratively trained with $\mathcal{L}_D^{\Omega}$.

Variant (2): As to the feedback triplet loss $\mathcal{L}_{FT}$, we design a variant feedback loss $\mathcal{L}_{FC}$ instead. We direct constraint the generated features $H_C^{\overline{\Omega}}$ in lesion-free areas to be similar to target features $H_T^{\overline{\Omega}}$.

The $\mathcal{L}_{FC}$ is defined as:

$$\mathcal{L}_{FC} = d_{tc} \quad (19)$$

where d_{tc} is defined as Eq. (10);

Therefore we have the final losses:

$$\mathcal{L}_2 = \mathcal{L}_G + \mathcal{L}_{NE} + \mathcal{L}_{FC} + \mathcal{L}_{CLS} \quad (20)$$

which are iteratively trained with $\mathcal{L}_D$. The L_G and L_D are the generator loss and the discriminator loss respectively, as we used in the competing loss in $\min_G \max_D \mathcal{L}_{AD}(G, D)$.

The experimental results of the two variants against our proposed method are shown in Table. IV. We can see that modifying either the adversarial loss $\mathcal{L}_{AD}(G, D)$ or the feedback triplet loss $\mathcal{L}_{FT}$ would lead to a descent performance.

TABLE III

AUC EVALUATION OF ABLATION STUDY ON (a) INBREAST DATASET FOR MASS CLASSIFICATION WITH ALEXNET; (b) INBREAST DATASET FOR MASS CLASSIFICATION WITH RESNET50; (c) INBREAST DATASET FOR MASS CLASSIFICATION WITH DENSENET; (d) INBREAST DATASET FOR MASS CLASSIFICATION WITH EFFICIENTNET; (e) INBREAST DATASET FOR MIXED-LESION CLASSIFICATION WITH RESNET50; (f) IN-HOUSE DATASET FOR MIXED-LESION CLASSIFICATION WITH ALEXNET

Bilateral	$\mathbf{L_{NE}}$	**Triplet Loss**	**AUC (Mass)**				**AUC (Mixed lesions)**	
			AlexNet [19]	ResNet [13]	DenseNet [17]	EfficientNet [37]	ResNet [13]	AlexNet [19]
×	×	×	0.820	0.827	0.822	0.830	0.780	0.697
SBF	×	×	0.862	0.858	0.855	0.861	0.807	0.721
TF-GAN	×	×	0.883	0.873	0.874	0.882	0.857	0.731
BF-GAN	×	×	0.860	0.842	0.846	0.861	0.849	0.720
AdaIN-GAN	×	×	0.886	0.873	0.877	0.888	0.858	0.734
AdaIN-GAN	✓	×	0.891	0.898	0.889	0.897	0.874	0.777
AdaIN-GAN	✓	Non-feedback	0.873	0.863	0.871	0.877	0.858	0.741
×	✓	Feedback	0.837	0.851	0.831	0.846	0.836	0.716
AdaIN-GAN	×	Feedback	0.905	0.902	0.903	0.902	0.884	0.771
AdaIN-GAN	✓	Feedback	**0.910**	**0.911**	**0.908**	**0.913**	**0.885**	**0.781**

TABLE IV

AUC EVALUATION ON (a) INBREAST DATASET FOR MASS CLASSIFICATION WITH ALEXNET; (b) INBREAST DATASET FOR MASS CLASSIFICATION WITH RESNET50; (c) INBREAST DATASET FOR MASS CLASSIFICATION WITH DENSENET; (d) INBREAST DATASET FOR MASS CLASSIFICATION WITH EFFICIENTNET; (e) INBREAST DATASET FOR MIXED-LESION CLASSIFICATION WITH RESNET50; (f) IN-HOUSE DATASET FOR MIXED-LESION CLASSIFICATION WITH ALEXNET

Experiment Setting	**AUC (Mass)**				**AUC (Mixed lesions)**	
Backbones	AlexNet [19]	ResNet [13]	DenseNet [17]	EfficientNet [37]	ResNet [13]	AlexNet [19]
Variant (1)	0.884	0.886	0.885	0.891	0.878	0.767
Variant (2)	0.860	0.863	0.862	0.867	0.850	0.739
Proposed Method	**0.910**	**0.911**	**0.908**	**0.913**	**0.885**	**0.781**

This can validate the effectiveness of these two losses in our method.

As we said that due to the pixel-to-pixel registration between bilateral images, we achieve counterfactual generation in feature level instead of image level. In practical experiments, we get 91.0% AUC of feature level which is higher than 90.6% of image level, verifying the performance of the feature generation. Moreover, the training speed of the former is more faster than the latter with 6.6 s/epoch v.s. 23.5 s/epoch.

G. Counterfactual Validation

As there are no ground truth images under counterfactual constraints, we adopt the visualization [4] and the FID measurement to validate the effectiveness of our counterfactual generation.

1) Counterfactual Visualization: We visualize the target features, reference features, and generated counterfactual features in Fig. 6. Since the three kinds of features are all with high dimension, we perform the max-pooling cross the channel dimension to generate the visualization heatmap for each of them. The heatmaps are shown in the last three columns respectively. Red indicates higher probabilities of being lesion areas and blue indicates lower probabilities of being lesion areas. As we can see, the activated lesion features (shown in red) in the target features marked by green rectangles disappear (shown in blue) in the corresponding areas of counterfactual features. Similar to the corresponding areas in the reference features, these areas are not highly responsive for lesions, which is consistent with the first statement of our Theorem III., *i.e.,* the distribution of counterfactual features in lesion areas should be equal to the distribution of reference features.[2]

And we can also see that the learned counterfactual features in other lesion-free areas are similar to the target features mostly. This is also consistent with the second statement of our Theorem III, *i.e.*, the counterfactual features and target features are not identical in pixel-to-pixel for the same reason as lesion areas. In summary, the visualized results show that the proposed method can effectively generate a lesion-free version of the target features, *i.e.,* counterfactual features.

We also visualize the predicted location of lesions during the iterative training process in Fig. 7, to further verify the effectiveness of our iterative optimization of CGN. As shown, with the process of iteration, the predicted location of lesions becomes more accurate.

2) FID Measurement: To further evaluate the effectiveness of the generated counterfactual features, we measure the feature distances in the INBreast by calculating the mean FID [15], which has been used for medical images [12], [24]. Specifically, the features from two distributions are fed to the Inception-V3 network respectively, of which the last pooling layer's output is taken as the final visual features for calculation. Denote μ_t, Σ_t, μ_r, Σ_r as the mean and covariance for the two distributions of features, then the FID d is computed by $d = ||\mu_t - \mu_r||_2^2 + Tr(\Sigma_t + \Sigma_r - 2(\Sigma_t \Sigma_r)^{1/2})$. The mean FID between the target and reference features is 56.15.

[2]Note that our result is given in the sense of distribution, it does not mean complete replacement from the reference features. Especially the glands in breasts are complicated, this means that the learned counterfactual features and the reference features in lesion areas cannot be identical in pixel-to-pixel.

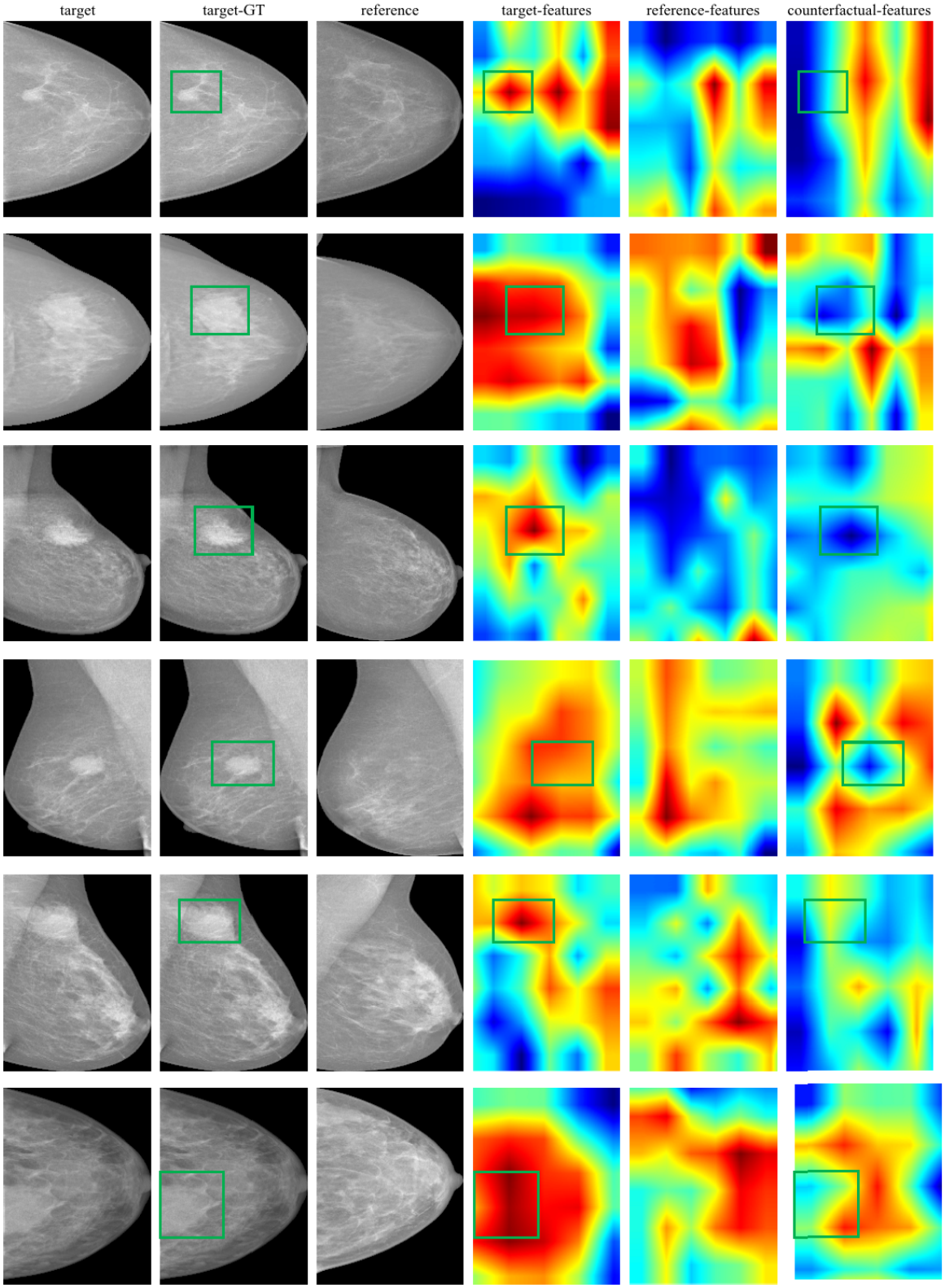


Fig. 6. **Visualization.** Left three columns: the target images, the target images with ground truth annotations which are marked by green rectangles on lesion areas, and reference images which are flipped horizontally for convenient comparison; Right three columns: feature maps of target images, feature maps of reference images, and feature maps of our generated counterfactual features. All visualized features are obtained by taking the maximum value of 256 channels. The green rectangles in each row mark the features in lesion areas before and after the counterfactual generation.

The counterfactual-reference mean FID is 27.04. The target-counterfactual mean FID is 25.42 while the one after removing the lesion areas from ground truth is 0.60. By comparing the four distances to each other, we find the learned counterfactual

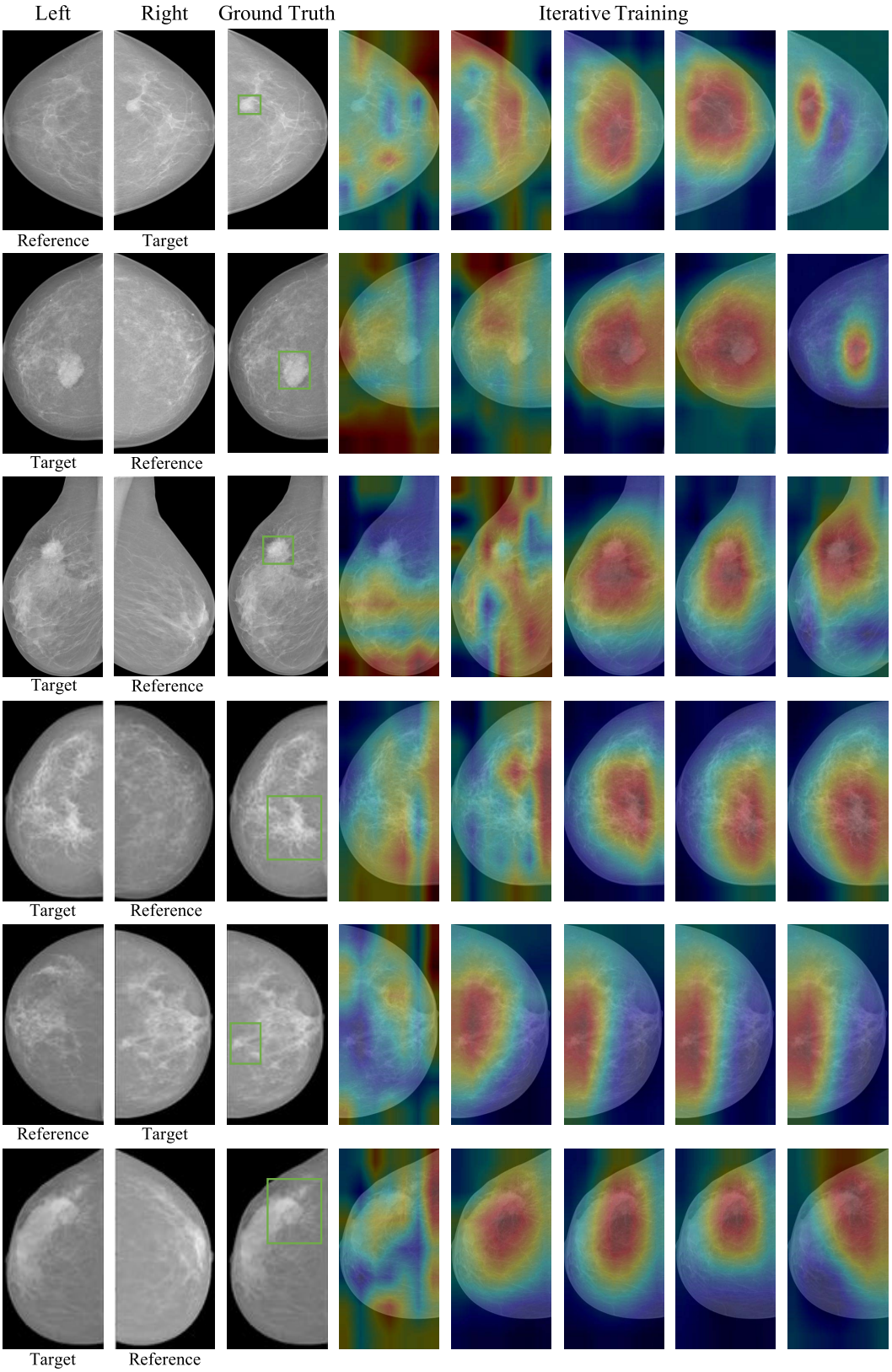


Fig. 7. **Iterative Training Process.** Left three columns: the images of the left side, the images of the right side, with being target or reference marked below, and the target images with ground truth annotations which are marked by green rectangles on lesion areas; Right five columns: the predicted location of lesions by CGN during training per ten epochs.

features contain both reference information in lesion areas (by FID of target-reference > FID of counterfactual-reference) and target information in lesion-free areas (by FID of reference-target > counterfactual-target).

TABLE V

AUC EVALUATION OF BILATERAL EXPERIMENTS ON (a) INBREAST DATASET FOR MASS MALIGNANCY CLASSIFICATION WITH ALEXNET; (b) INBREAST DATASET FOR MASS MALIGNANCY CLASSIFICATION WITH RESNET50; (c) INBREAST DATASET FOR MASS CLASSIFICATION WITH DENSENET; (d) INBREAST DATASET FOR MASS CLASSIFICATION WITH EFFICIENTNET; (e) INBREAST DATASET FOR MIXED-LESION MALIGNANCY CLASSIFICATION WITH RESNET50; (f) IN-HOUSE DATASET FOR MIXED-LESION MALIGNANCY CLASSIFICATION WITH ALEXNET

Experiment Setting	AUC (Mass)				AUC (Mixed lesions)	
Backbones	AlexNet [19]	ResNet [13]	DenseNet [17]	EfficientNet [37]	ResNet [13]	AlexNet [19]
SBF	0.862	0.858	0.855	0.861	0.807	0.721
GF	0.865	0.862	0.858	0.864	0.812	0.726
SFF	0.864	0.862	0.861	0.866	0.813	0.724
Proposed Method	**0.910**	**0.911**	**0.908**	**0.913**	**0.885**	**0.781**

H. Bilateral Distribution Verification

In this section, we verify the correctness of our symmetric prior assumption that motivates our proposed framework. Specifically, we choose 1,000 unhealthy bilateral images, each of which contains at least one lesion from the in-house dataset. Then for comparison, we choose another 1,000 healthy bilateral images.[3] We respectively calculate FID to measure the distribution similarities for the healthy set (*i.e.*, D^H) and the unhealthy set (*i.e.*, D^U). Then we conduct the one-tailed T-test with the null hypothesis H_0 and alternative hypothesis H_1 defined as:

$$H_0 : \mu(D^H) >= \mu(D^U) \quad H_1 : \mu(D^H) < \mu(D^U).$$

We obtain a p-value of $0.014 < 0.05$, which means that we can reject the original hypothesis H_0 at a 95% significance level. This result provides an evidence for us to reject H_0, i.e., the bilateral distribution distance of unhealthy cases is larger than healthy cases significantly. This result can be regarded as a manifestation of our symmetric prior assumption.

V. CONCLUSIONS

In this paper, we propose a novel approach called bilateral asymmetry guided Counterfactual Generating Network (CGN) to improve the mammogram classification performance. The proposed method performs the counterfactual generation by exploiting the symmetric prior effectively. Experimental results indicate that the proposed CGN achieves state-of-the-art results in both public and in-house datasets. Our work can be referred as the showcase of exploiting symmetric prior, which widely holds in many human organs, *e.g.,* brains, eyes, skeletal structures, and kidneys. Therefore, we believe that the generalization ability of our method on corresponding medical imaging problems, the efforts of which will be left in future work.

APPENDIX A
PROOF OF THEOREM 1

Lemma 2: If the causal graph $\mathcal{G}$ satisfies that the common factor C influences the bilateral variables simultaneously, then,

$$\begin{aligned} f_{Y_T}(C,\cdot) &= f_{Y_R}(C,\cdot) \\ f_{H_T}(C,\cdot) &= f_{H_R}(C,\cdot) \\ f_{X_T}(C,\cdot) &= f_{X_R}(C,\cdot) \end{aligned} \tag{21}$$

[3]We do not use the public INBreast dataset since there are too few healthy couples for the FID results to be statistically significant

Lemma 2 shows that the causal factor C influences the bilateral mammograms in equal function relationship.

Proof of Theorem 1:

Proof of Eq. (2)*:*

$$\begin{aligned} &P({H_T}_{(Z_T^\Omega=0)}{}^{\Omega} = h | H_T^\Omega = h_t, Z_T^\Omega = 1) \\ &= \int_c P({H_T}_{(Z_T^\Omega=0)}{}^{\Omega} = h | C = c) P(C = c | H_T^\Omega = h_t, Z_T^\Omega = 1) dc \\ &= \int_c P(H_T^\Omega = h | C = c, Z_T^\Omega = 0) P(c | H_T^\Omega = h, Z_T^\Omega = 1) dc \\ &= \int_c P(H_R^\Omega = h | C = c, Z_R^\Omega = 0) P(c | H_T^\Omega = h, Z_T^\Omega = 1) dc \\ &= P({H_R}_{(Z_R^\Omega=0)}{}^{\Omega} = h | H_T^\Omega = h_t, Z_T^\Omega = 1), \\ &= P(H_R^\Omega = h_r | H_T^\Omega = h_t, Z_T^\Omega = 1), \end{aligned} \tag{22}$$

where the first equation is due to that the c is the only parent node of ${H_T}_{(Z_T^\Omega=0)}{}^{\Omega}$; the second equation is according to Markov condition that ${H_T}_{(Z_T^\Omega=0)}{}^{\Omega} | C = H_T | C, Z_T^\Omega$, the third equation is due to the symmetric prior.

Proof of Eq. (3)*:* Since in the lesion-free areas, there are $Z_T^{\overline{\Omega}} = 0$, the probabilities are derived by the actual hidden features $H_T = h_t$ directly, i.e.,

$$\begin{aligned} &P({H_T}_{(Z_T^{\overline{\Omega}}=0)}{}^{\overline{\Omega}} = h | H_T^{\overline{\Omega}} = h_t, Z_T^{\overline{\Omega}} = 0) \\ &\qquad = P(H_T^{\overline{\Omega}} = h_t | H_T^{\overline{\Omega}} = h_t, Z_T^{\overline{\Omega}} = 0) \end{aligned} \tag{23}$$

□

APPENDIX B
BILATERAL ANALYSIS

To validate the superiority of our method over others in identifying lesion regions for prediction based on bilateral images, we compare with some bilateral fusion mechanisms in some lesion detection methods. For our setting that is without any ROI annotations and with image-level benign/malignant ground truth only, we adapt their fusion mechanisms of bilateral images for malignant prediction. We apply the same classifier with ours, which takes the fused bilateral features extracted by these methods as input. The compared bilateral fusion mechanisms are listed below:

SBF: **Simple Bilateral Fusion**, as mentioned in Section IV-F.

GF: **Gated Fusion**. Assign a gate to each bilateral region feature to obtain the weighted bilateral features, with the weight calculated via the gate operator [22].

SFF: **Simple Four-view Fusion**. Ensemble cross-view and contralateral-view by simple information fusion [39].

Both **SFF** and **GF** can be seen as variants of **SBF**. Specifically, one branch of the model (view-wise fusion) in the paper [39] and the gated fusion performed to proposals in the 2nd paper [22] can be applied as bilateral feature fusion like one of our ablation experiments in Tab. III (**SBF**, Simple Bilateral Fusion). Due to better feature integration or leveraging more cross-view information, the **GF** and **SFF** slightly outperform **SBF**, as shown in Table. V. But both of them share the similar disadvantage with **SBF**: even for healthy breasts, bilateral mammograms are only roughly symmetric but not pixel-to-pixel, the similarity of bilateral features cannot be guaranteed. While our method uses the symmetry prior by healthy generation with an improved GAN. Therefore, our method suffers less from problems and leads to better results.

REFERENCES

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3D point clouds," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 40–49.

[2] R. Alterson and D. B. Plewes, "Bilateral symmetry analysis of breast MRI," *Phys. Med. Biol.*, vol. 48, no. 20, pp. 3431–3443, Oct. 2003.

[3] J. Amores and P. Radeva, "Retrieval of IVUS images using contextual information and elastic matching," *Int. J. Intell. Syst.*, vol. 20, no. 5, pp. 541–559, 2005.

[4] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, "Counterfactuals uncover the modular structure of deep generative models," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Apr. 2020.

[5] J. W. Byng *et al.*, "Symmetry of projection in the quantitative analysis of mammographic images," *Eur. J. Cancer Prevention*, vol. 5, no. 5, pp. 319–327, Oct. 1996.

[6] V. Chernozhukov, "Inference on counterfactual distributions," *Econometrica*, vol. 81, no. 6, pp. 2205–2268, 2013.

[7] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.

[8] N. Dhungel, G. Carneiro, and A. P. Bradley, "The automated learning of deep features for breast mass classification from mammograms," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 106–114.

[9] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10705–10714.

[10] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[11] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[12] C. Haarburger *et al.*, "Multiparametric magnetic resonance image synthesis using generative adversarial networks," in *Proc. VCBM*, 2019, pp. 11–15.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[16] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, "Mask-ShadowGAN: Learning to remove shadows from unpaired data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2472–2481.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[18] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[20] H. Lee *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomed. Eng.*, vol. 3, no. 3, pp. 173–182, 2019.

[21] H. Li, K. R. Mendel, L. Lan, D. Sheth, and M. L. Giger, "Digital mammography in breast cancer: Additive value of radiomics of breast parenchyma," *Radiology*, vol. 291, no. 1, pp. 15–20, 2019.

[22] Y. Liu *et al.*, "From unilateral to bilateral learning: Detecting mammogram masses with contrasted bilateral network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 477–485.

[23] W. Lotter, G. Sorensen, and D. Cox, "A multi-scale CNN and curriculum learning strategy for mammogram classification," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 169–177.

[24] I. Malkiel, S. Ahn, V. Taviani, A. Menini, L. Wolf, and C. J. Hardy, "Conditional WGANs with adaptive gradient balancing for sparse MRI reconstruction," 2019, *arXiv:1905.00985*. [Online]. Available: http://arxiv.org/abs/1905.00985

[25] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.

[26] O. Nizan and A. Tal, "Breaking the cycle–colleagues are all you need," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7860–7869.

[27] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[28] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[29] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[30] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Sci. Rep.*, vol. 8, no. 1, p. 4165, Dec. 2018.

[31] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 146–157.

[32] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[33] E. Sickles, C. J. D'Orsi, and L. W. Bassett, *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*, vol. 2014. Reston, VA, USA: American College of Radiology, 2013, pp. 37–78.

[34] M. M. R. Siddiquee *et al.*, "Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 191–200.

[35] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA, Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[36] S.-C. Tai, Z.-S. Chen, and W.-T. Tsai, "An automatic mass detection system in mammograms based on complex texture features," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 2, pp. 618–627, Mar. 2013.

[37] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[38] C.-R. Wang, F. Zhang, Y. Yu, and Y. Wang, "BR-GAN: Bilateral residual generating adversarial network for mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 657–666.

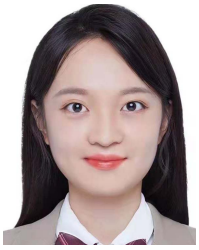
[39] J. Wei *et al.*, "Computer-aided detection of breast masses: Four-view strategy for screening mammography," *Med. Phys.*, vol. 38, no. 4, pp. 1867–1876, 2011.

[40] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling gans for data augmentation in mammogram classification," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Cham, Switzerland: Springer, 2018, pp. 98–106.

[41] N. Wu *et al.*, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020.

[42] Y. Xie and D. Richmond, "Pre-training on grayscale imagenet improves medical image classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018.
[43] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2017.
[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[46] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 603–611.

Churan Wang received the bachelor's degree in exploration technology and engineering from China University of Petroleum, Beijing, in 2017. She is currently pursuing the Ph.D. degree in data science (computer science) with Peking University. Her research interests include computer vision and medical image analysis.



Jing Li received the bachelor's degree from the School of Mathematics and Statistics, Wuhan University, in 2017. She is currently pursuing the Ph.D. degree in computer science with Peking University. Her research interests include computer vision, causal inference, reinforcement learning, and their applications.

Fandong Zhang received the Ph.D. degree in EECS from Peking University. He currently holds a postdoctoral position with the Academy for Advanced Interdisciplinary Studies, Peking University. His research interests include medical image analysis, computer vision, and biometrics. Recently, he is working on automatic analysis of breast medical imaging, including mammography, MRI and ultrasound.

Xinwei Sun received the Ph.D. degree from the School of Mathematical Sciences, Peking University, in 2018. He is currently an Incoming Assistant Professor with Peking University. His research interests mainly focus on statistical machine learning, causal inference, with their applications on medical imaging, computer vision, and few-shot learning.

Hao Dong (Member, IEEE) received the B.Eng. degree (Hons.) from the University of Central Lancashire, the M.Sc. (Hons.) degree from Imperial College London, U.K., and the Ph.D. degree from Imperial College London in fall 2019, under the supervision of Yike Guo. He is currently an Assistant Professor with Peking University studying AI, vision, and machine learning. He is a member of Peng Cheng Laboratory. He founded a startup for brain–computer interface with Yike Guo from 2012 to 2014.

Yizhou Yu (Fellow, IEEE) received the Ph.D. degree from the University of California at Berkeley, in 2000. He is currently a Professor with The University of Hong Kong. He was a Faculty Member at the University of Illinois at Urbana–Champaign for twelve years. He is also the Chief Scientist at Deepwise Healthcare. His current research interests include computer vision, deep learning, AI in medicine, computational visual media, and geometric computing. He was a recipient of the 2002 US National Science Foundation CAREER Award and the ACCV 2018 Best Application Paper Award. He has served on the Editorial Board of *IET Computer Vision*, *The Visual Computer*, and IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS. He has also served on the program committee of many leading international conferences, including CVPR, ICCV, and SIGGRAPH.

Yizhou Wang received the bachelor's degree in electrical engineering from Tsinghua University in 1996 and the Ph.D. degree in computer science from the University of California at Los Angeles (UCLA) in 2005. He joined the Xerox Palo Alto Research Center (Xerox PARC) as a Research Staff from 2005 to 2007. He is currently a Boya Professor with the Computer Science Department and the Vice Director of the Center on Frontiers of Computing Studies, Peking University. He was granted the National Natural Science Fund (NSFC) for Distinguished Young Scholars. His research interests include computational vision, statistical modeling and learning, medical image analysis, and computational arts.