Cascaded Generative and Discriminative Learning for Microcalcification Detection in Breast Mammograms

Fandong Zhang[1,5*] Ling Luo[2,3*] Xinwei Sun[2,5] Zhen Zhou[2,5] Xiuli Li[2,5]
Yizhou Yu[2] Yizhou Wang[2,4,5]
[1]Center for Data Science, Peking University [2]Deepwise AI Lab
[3]School of Cyberspace Security, Beijing University of Posts and Telecommunications
[4]Computer Science Department, School of EECS, Peking University [5]Peng Cheng Laboratory
{fd.zhang, z.zhou, yizhou.wang}@pku.edu.cn {luoling, sunxinwei, lixiuli, yuyizhou}@deepwise.com

Abstract

Accurate microcalcification (μC) detection is of great importance due to its high proportion in early breast cancers. Most of the previous μC detection methods belong to discriminative models, where classifiers are exploited to distinguish μCs from other backgrounds. However, it is still challenging for these methods to tell the μCs from amounts of normal tissues because they are too tiny (at most 14 pixels). Generative methods can precisely model the normal tissues and regard the abnormal ones as outliers, while they fail to further distinguish the μCs from other anomalies, i.e. vessel calcifications. In this paper, we propose a hybrid approach by taking advantages of both generative and discriminative models. Firstly, a generative model named Anomaly Separation Network (ASN) is used to generate candidate μCs. ASN contains two major components. A deep convolutional encoder-decoder network is built to learn the image reconstruction mapping and a t-test loss function is designed to separate the distributions of the reconstruction residuals of μCs from normal tissues. Secondly, a discriminative model is cascaded to tell the μCs from the false positives. Finally, to verify the effectiveness of our method, we conduct experiments on both public and in-house datasets, which demonstrates that our approach outperforms previous state-of-the-art methods.

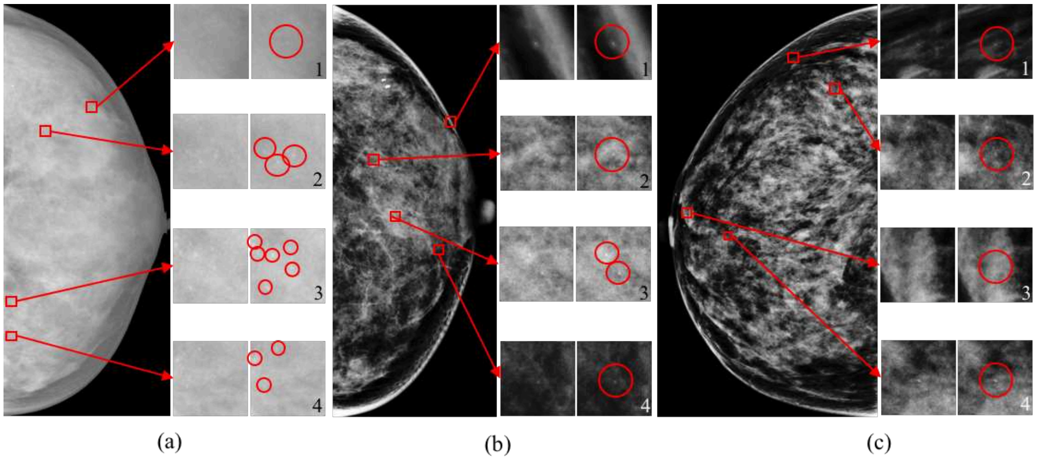
1. Introduction

Breast cancer is the most common cancer among women worldwide [28]. To discover it at the early state, breast screening is necessarily applied [27]. Among the signs of early breast cancers, the microcalcifications (μCs) belongs to one of the most common kinds [2]. To analyze them, the mammogram images are widely used. As shown in Fig.1, μCs are tiny and vary in brightness, contrast, shape with diverse surroundings. It is obviously difficult and time consuming for radiologists to detect them one by one. Therefore, an automatical μCs detection methods with high accuracy in mammography images is of great importance.

To achieve this goal, different methods are proposed, among which most are discriminative models, *i.e.*, classification models. Usually, various features, such as harr-like features [31, 3], shape and texture features [14] and deep features [4, 23] are extracted from images to train a binary classifier that can tell μC pixels from the normal ones. However, these methods suffer from extremely imbalanced samples. The reason lies in that μCs are commonly too tiny, generally smaller than 14 pixels in mammogram images, and the vast majority of such image regions are normal tissues. Therefore, it is challenging to extract efficient features for such small objects and also lead to terrible distribution of μC and other tissues. In our experiments, the ratio between positive (μCs) samples and negative (normal tissues) samples is around 4×10^3.

To address the aforementioned problem, we try to firstly distinguish normal pixels and abnormal ones, while μCs belongs to abnormal regions. In this way, we can reduce a lot of negative samples that exist in discriminative models. To this end, we rethink the μC detection task from the point of image reconstruction. The normal samples are those regular backgrounds and there are a huge mount of these regions. Hence it is supposed to be not hard to find a dictionary to reconstruct these normal samples. On the contrary, μCs are irregular, rare and hard to be reconstructed. Therefore, it is natural to learn a reconstruction model that normal samples can be well reconstructed while μCs not. In this paper, we design an image reconstruction network, which is modeled as a deep convolutional encoder-decoder network, providing the reconstruction function with powerful learning abil-

[1]Authors contributed equally.



(a) (b) (c)

Figure 1. An illustration of μCs in mammogram images. (a) is sampled from INBreast dataset [19], (b) and (c) are from our in-house dataset. Three columns in each sub-figure represent the raw mammogram image, patches zooming in μCs, and the corresponding circled μCs, respectively. Most μCs are within 1mm, which is about 14 pixels in a mammogram image with 70um pixel spacing. Images from INBreast and in-house dataset have very different gray-level histograms due to different data resources. μCs from (b) and (c) show diversity in brightness, contrast, shape and surroundings.

ity. Moreover, to extract informative features for such tiny objects, U-Net [24] is exploited as the backbone network.

To further improve the performance of such reconstruction procesure, a novel t-test loss is designed to drive the distribution of the residual of μCs away from that of normal regions. The proposed t-test loss is inspired by the two-sample t-test, which is a classical hypothesis test method. Here, we alternate it into a data-driven loss function. Specifically, we regard the reconstruction residuals of positive (μCs) and negative (normal tissues) pixels as two independent random variables learned by the reconstruction network. Instead of determining whether such two distributions are different or not, our t-test loss forces the reconstruction network to constrain these two distributions differently as much as possible. Since the normal tissues are easy to reconstruct while μCs are not, we minimize the reconstruction residuals of negative pixels and implement a hard thresholding to constrain the positive ones to be large than a pre-set threshold. So far, we have explained all necessary components of the proposed generative module, which is called Anomaly Separation Network (ASN).

After the reconstruction, abnormal regions are obtained, which contain both candidates μCs and other kinds of calcifications, such as vessel calcifications, rod-like calcifications, etc. Although they all belong to calcifications, we observe that they are quite different from each other in shape. Benefited from such a property, we build a discriminative model, *i.e.*, a deep binary classification network, to classify μCs and others. This discriminative model is designed to implement the False Positive Reduction (FPR).

To verify the effectiveness of the proposed method, we implement evaluation on both public dataset INBreast [19] and our in-house dataset. We achieve a recall of 78.35% at 5 false positive per image (simplified as R@5) and a R@10 of 85.96% on InBreast; a R@5 of 90.71% and a R@10 of 92.24% on in-house dataset, which outperforms previous state-of-the-art methods.

To summary, our contributions are mainly three-fold: 1) To solve the imbalanced problem that previous discriminative models suffer, we propose a generative model to distinguish the normal regions from abnormal ones where the candidates μCs lie in. Moreover, U-Net is applied to extract informative features for such tiny objects. 2) To further enhance the performance, a novel t-test loss is designed to enlarge the distribution diversity between normal regions and abnormal ones. 3) The proposed ASN achieves the best performance on both public and in-house datasets, compared with previous state-of-the-art methods.

2. Related Works

2.1. Microcalcification Detection

Most existing μC detection approaches can be coarsely classified into two categories: image processing based and learning based. The first category is mainly based on the fact that μCs are commonly brighter with higher frequency than their surrounding tissues. Mammogram images are first enhanced with wavelet transform [1, 14] and then hessian matrix response [20], morphological filtering [1] are applied to identify μCs. However, such methods are easily affected by dense tissues and also suffer from the large mount of false positives.

The second category is based on supervised learning. Effective binary classifiers can be trained to tell the μCs from normal tissues. Khalaf et al. [14] extract several shape and texture descriptiors, and apply Students t-test and SVM with RBF kernels for feature selection and training. Harr-like features [31] are used in [3], where a set of cascaded classifiers are trained to cope with the class imbalance problem. Cai et al. [4] apply CNN to learn deep features for classification. The network is trained on proposals generated by thresholding on band-pass filtered images.

2.2. Image Reconstruction

Image reconstruction is the problem of reconstructing original image which might be noisy and blurred [22]. One typically applied method is sparse dictionary learning which aims at learning the sparse linear representation of the elements that altogether compose dictionary. Implemented with L_1 regularization, the learned representation is robust to occlusion. For example, a classification algorithm based on sparse coding was proposed [33] to successfully handle occlusion and corruption uniformly and robustly in recognition of face images.

To benefit from the powerful representation ability of CNN, Turchenko et al. [30] present a deep convolutional auto-encoder to achieve dimension reduction, clustering and image reconstruction. Kingma and Welling [16] design the auto-encoding variational bayes algorithm which allows us to perform very efficient approximate posterior inference that can also be used for a host of tasks such as recognition, denoising, representation and visualization purposes. Johnson et al. [13] propose the use of perceptual loss functions for training feed-forward networks for image transformation tasks. Goodfellow et al. [8] build a new framework for estimating generative models via an adversarial process, which can be widely used for image generation.

2.3. Two-sample T-test

The two-sample T-test is a statistical hypothesis testing method to determine whether the two sets of data are significantly different from each other. Assuming that the samples in each group are independent and identically distributed from normal distribution[1], then the computed $t\text{-}statistics$ follows a Student's t-distribution under null hypothesis that two groups are not differently distributed. By product, the corresponding p-value, which measures the probability that the null hypothesis holds, can also be calculated. Hence, one can reject the null hypothesis when the p-value is less than pre-defined threshold level α, which means that the two groups of data are thought to be differently distributed.

2.4. Anomaly Detection

Anomaly detection, also referred as outlier detection, is the problem of identification of patterns in data that do not conform to expected behavior [5]. Such non-conforming patterns, i.e. outliers are generally defined as the anomalies, rare events or aberrant data suspected to be generated from a different mechanism that is deviate markedly from the most common or expected pattern [7]. The detection of outliers may provide us important information, e.g. credit card fraud, medical problems in clinical trials. Moreover, the existence of such outliers may result in the instability in estimation, inference, and model selection, etc. Hence, the outlier identification is a critical task to obtain robust parameter estimation and detecting anomalies given new data [12]. In our paper, we consider the μCs as outliers since the number of positive image pixels are rare and differently distributed compared to regular negative ones.

Various methods have been proposed for outlier detection, including univariate models [17] and multivariate models [32, 12, 25]. For unsupervised outlier detection where the anomalies are unlabeled, one can typically apply robust regression with Hubers loss [12], which minimizes square loss for normal data and absolute loss for abnormal ones. It has been proved in [26] that this scheme is equivalent to a LASSO problem, which translates the detection of anomalies into a model selection problem.

3. Methodology

As shown in Fig. 2, our system mainly consists of two cascaded modules: Anomaly Separation Network (ASN) and False Positive Reduction (FPR) model. The outputs of ASN are directly fed into FPR, which predicts the final results. In this section, we will demonstrate ASN and FPR separately.

3.1. Anomaly Separation Network

ASN includes two core components: deep reconstruction network and reconstruction residual learning with t-test loss. During training process, mammogram images are cut into patches and sent into a deep reconstruction network. Then we apply the t-test loss on the reconstruction residuals, which encourages the residuals of normal pixels to be small, while μCs pixels to be relatively large. During testing process, for each whole mammogram, we calculate the reconstruction residual map and predict points and scores based on it. In this section, we will first demonstrate our reconstruction network and reconstruction residual learning, then explain the connection between t-test loss and Huber's loss [12] which is widely used for anomaly detection.

3.1.1 Deep Reconstruction Network

We design a deep reconstruction network to provide a learnable reconstruction function. Deep ConvNet has been proved robust and effective for many image tasks. To take

[1] The assumption of normal distribution can be relaxed according to central limit theorem

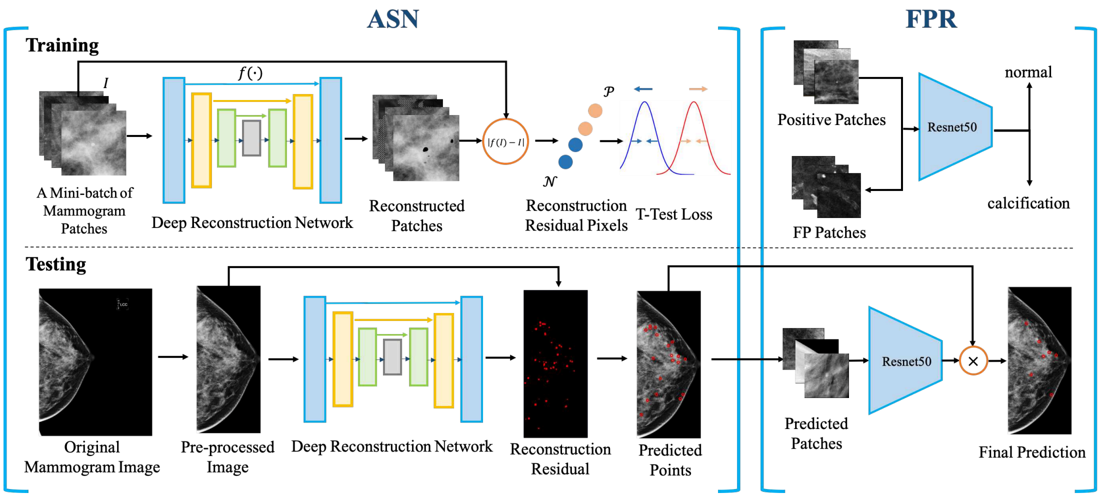


Figure 2. The pipeline of the proposed method. There are two cascaded models: Anomaly Separation Network (ASN) and False Positive Reduction (FPR) model. The outputs of ASN are directly fed into FPR, which predicts the final results. For ASN, during training process, the U-Net based reconstruction network is trained with mammogram patches. T-test loss is applied on the reconstruction residual pixels to drive the normal residuals and μC residuals away from each other. During testing process, given a mammogram image, after pre-processing, the reconstruction residual is computed, which can generate predicted points (show in red circles). FPR model is a ResNet50, trained by hard negatives of ASN and positives of ground truth. The final prediction is a fusion of both models by product in score level.

the great representation ability, we use a U-Net [24] for pixelwise reconstruction. Our U-Net consists of 3 downsample stages and 3 upsample stages with skip connections. Each stage includes 3 convolution layers.

We design such a network for three reasons. Firstly, the downsampling operations can lead to effective receptive field size, which is advantageous for the reconstruction of each pixel and the coherence of reconstructed image. Secondly, sizes of μCs are within 14 pixels. Therefore, we only downsample the image by the factor of 8 to avoid too much information loss for reconstruction. Thirdly, the skip connections can keep low-level information, which is necessary for accurate localization.

3.1.2 Reconstruction Residual Learning

Let $f(\cdot)$ denote the reconstruction network function. Given an image I, the reconstruction residual value is,

$$\begin{aligned} r(I) &= |f(\Theta; I) - I| \\ &= \sum_{\mathcal{P}\in I} |f(\Theta;\mathcal{P}) - \mathcal{P}| \ \Big(f(\Theta;\mathcal{P}) \triangleq f(\Theta; I)[\mathcal{P}]\Big) \end{aligned} \tag{1}$$

where $\mathcal{P}$ denotes the pixel and Θ denotes the parameters in the reconstruction network. The reconstruction residual of positive and negative pixels are desired to have different distributions. Therefore, we propose the t-test loss. In the following sections, we will firstly review the two-sample t-test, and then demonstrate effectiveness of the t-test loss.

Two-Sample T-test Given two groups of samples $x_1, ..., x_{N_x} \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ and $y_1, ..., y_{N_y} \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$, to test whether $\mu_x > \mu_y$ or not, we build null hypothesis (H_0) and alternative hypothesis (H_1) [6] as,

$$H_0 : \mu_x <= \mu_y \quad H_1 : \mu_x > \mu_y \tag{2}$$

And a *t-statistics* is generated using the following formula:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{N_x} + \frac{S_y^2}{N_y}}} \tag{3}$$

where $\bar{\cdot}$ denotes the mean value of a group of samples, S_x and S_y are sample variances of x and y, respectively. We choose to accept H_1 (reject H_0) if $t \geqslant t_{\nu,\alpha}$ where $t_{\nu,\alpha}$ is the critical value at significant level α of the Student's t-distribution with degrees of freedom ν, i.e. $P(t \geqslant t_{\nu,\alpha}) = \alpha$, where

$$\nu = \begin{cases} \dfrac{\left(\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}\right)^2}{\frac{\sigma_x^2}{N_x}\frac{1}{N_x-1} + \frac{\sigma_y^2}{N_y}\frac{1}{N_y-1}}, & \text{if } \sigma_x \neq \sigma_y \\ N_x + N_y - 2, & \text{if } \sigma_x = \sigma_y \end{cases} \tag{4}$$

In real applications, the $\{x_i\}_{i=1,...,N_x}$ and $\{y_j\}_{j=1,...,N_y}$ may not ideally satisfy normal distribution. However, according to the central limit theorem, the $\bar{x}$ and $\bar{y}$ are approximate to normal distribution when N_x and N_y are large enough, in which the two-sample t-test can also be applied.

Since the normal tissues are regular and calcifications are not hence scattered distributed, we estimate the degree of freedom as

$$\widetilde{\nu} = \frac{\left(\frac{S_x^2}{N_x} + \frac{S_y^2}{N_y}\right)^2}{\frac{S_x^2}{N_x}\frac{1}{N_x-1} + \frac{S_y^2}{N_y}\frac{1}{N_y-1}} \tag{5}$$

T-test Loss Given independent negative and positive samples, we use Eq. 1 to compute the residual value of reconstruction, denoted as $\{r_p^i(\Theta)\}_{i=1,\dots,N_p}$ and $\{r_n^i(\Theta)\}_{i=1,\dots,N_n}$. In the rest of the paper, we denote the above residual values as $\{r_p^i\}_{i=1,\dots,N_p}$ and $\{r_n^i\}_{i=1,\dots,N_n}$ for simplicity. We then proposing the following t-test loss,

$$\mathcal{L} = \max(\beta - \bar{r_p}, 0) + \bar{r_n} + \lambda_p S_{r_p}^2 + \lambda_n S_{r_n}^2 \qquad (6)$$

where the threshold hyper-parameter β denotes the margin between the means of positive and negative residual distributions; λ_p and λ_n are regularization hyper-parameters.

Minimizing such t-test loss can be viewed as maximizing the $t\text{-}statistics$ defined in Eq. 2, which is commonly used to determine whether two groups of data are different from each other. Our goal instead, is accurate classification, i.e. ability to discriminate the μCs from negative image pixels in a supervised way. To achieve this goal, we in turn propose to drive reconstruction of labeled positive pixels away from negative ones by minimizing $\mathcal{L}$.

In more details, note that $\max(\beta - \bar{r_p}, 0) + \bar{r_n}$ hopes the reconstruction parameter (Θ) to well fit the negative pixels while leave the reconstruction of positive pixels with a large margin. In another way, Θ is trained to learn the negative pixels and also the remaining pixels except the μCs in the positive patch. Therefore, for positive pixels in the test data, the Θ can reconstruct with big margin. In such way, it can be successfully predicted as positive label and the corresponding residue can be regarded as the μC.

Besides such part in Eq. 6, we additionally regularize S_x and S_y. Without such regularization, the estimation of Θ tends to be unstable. That's because large values of S_x and S_y can make $r_{t=n,p}^i$ easy to be either small or large since they tend to be distributed in a widely spread way.

In contrast, the estimation error loss $\min \bar{r_n} + \bar{r_p}$ may suffer from the model collapse problem that the learned mapping function tends to be identity. Therefore, they are unable to model the underlying structure of positive image pixels and hence can not be generalized to detect μC in the test phase. Moreover, compared with the estimation error loss, our loss is more task-driven since it's agreed with the rule of outlier detection in the test phase, i.e. the patch i is detected as outlier if $r^i > \beta$.

In addition, the estimated Θ, which is supervised (the μC are labeled) to model the residual values of negative (positive) samples less (larger) than the threshold parameter β, can be directly used to detect μC in the test phase. Hence, the t-test loss can be incorporated into the whole end-to-end procedure, which is illustrated in Fig. 2.

3.1.3 Connection to Huber's loss

We claim that the proposed t-test loss, i.e. Eq 6 can be viewed as the variation of robust regression with Huber's loss [12], which is an unsupervised outlier detection method. In more details, note that the Huber's loss in our scenario can be written as:

$$\mathcal{L}_{Huber}(\Theta) = \sum_{i=1}^{N} \rho_\beta(\mathcal{P}^i, f(\Theta; \mathcal{P}^i)) \qquad (7)$$

where N denotes the number of patches in training set and

$$\rho_\beta(\mathcal{P}^i, f(\Theta; \mathcal{P}^i)) = \begin{cases} \frac{1}{2}\left(\mathcal{P}^i - f(\Theta; \mathcal{P}^i)\right)^2, & \left|\mathcal{P}^i - f(\Theta; \mathcal{P}^i)\right| \leqslant \beta \\ \beta\left(\left|\mathcal{P}^i - f(\Theta; \mathcal{P}^i)\right| - \frac{1}{2}\beta\right), & \text{otherwise} \end{cases} \qquad (8)$$

Eq. 7 is the combination of square loss (mean unbiased estimators) and absolute loss (median unbiased estimators). It has been proved in [26] that the minimization of Eq. 7 is equivalent to

$$\min_{\Theta} \sum_{i=1}^{N_n} \frac{1}{2}\left(\mathcal{P}_n^i - f(\Theta; \mathcal{P}_n^i) - \gamma^i\right)^2 + \sum_{i=1}^{N_p} \frac{1}{2}\left(\mathcal{P}_p^i - f(\Theta; \mathcal{P}_p^i) - \gamma^i\right)^2 + \lambda\|\gamma\|_1, \quad (9)$$

i is outlier, i.e. $\gamma^i \neq 0$ if and only if $\left|\mathcal{P}^i - f(\Theta; \mathcal{P}^i)\right| > \beta$. Here the outliers are unlabeled and they can be regarded as the elements with non-zero value of γ.

In our experiments, the outliers (positive image pixels $\mathcal{P}_p$) in the training data are labeled. Hence, to remove the deviating effect of such outliers, we in turn propose to constrain such outliers to satisfy the definition in Huber's loss, i.e. $\left|\mathcal{P}_p^i - f(\Theta; \mathcal{P}_p^i)\right| > \beta$ for each i. Combined with the absolute loss for $\mathcal{P}_n$, the total loss can be correspondingly designed as:

$$\widetilde{\mathcal{L}} = \max(\beta - \bar{r_p}, 0) + \bar{r_n}. \qquad (10)$$

The $\bar{r_p}$ is expected to be larger than the threshold parameter β, which is a relaxation of the constraint in Huber's loss for more robustness and better generalization. By removing the outliers' effect, a robust estimation of Θ can be achieved, which in turn can lead to accurate detection of outliers in the test data. Besides, we additionally regularize variances to prevent unstable parameter estimation, as mentioned earlier.

3.1.4 Setting Hyper-parameters

The threshold parameter $\beta > 0$ is inversely proportional to significance level α. From the view of outlier detection, it is the trade-off between "masking effect" and "swamping effect" [12]. Too small β may lead to incorrectly identifying negative pixel as outlier, i.e. swamping effect; while too

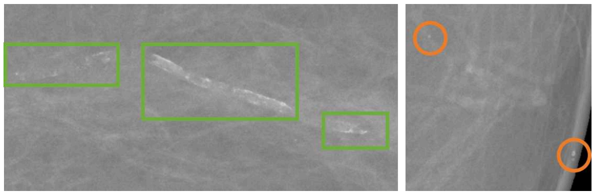


Figure 3. Examples of comparison vessel calcifications (left, marking in green rectangles) with μCs (right, marking in orange circles).

Table 1. Evaluations on INBreast dataset (%).

Method	R@1	R@5	R@10	R@15	R@20
FPN FRCN	**39.72**	71.47	72.48	72.48	72.48
U-Net w FPR	29.45	77.61	82.84	83.67	84.50
Proposed	36.70	**78.35**	**85.96**	**88.26**	**88.90**

large value may result in missing some outliers, i.e. masking effect.

Larger $\lambda_{s=n,p}$ implies more regularization on variances. Here, we incorporate the heterogenous regularization ($\lambda_p \neq \lambda_n$) into our loss, which means that the variances are different. In our experiments, the μC is irregular hence the reconstruction may vary a lot. Hence, it's reasonable to implement larger regularization on S_p to prevent S_p from being too large. As what'll be shown in the experiment section, the best prediction results are given when $\lambda_p > \lambda_n$.

3.2. False Positive Reduction

The proposed ASN can reconstruct normal tissues well and regard μCs as anomalies. However, there are kinds of calcifications in breast mammograms. As is shown in Fig.3, the green rectangles in the left patch are vessel calcifications, which can be considered as lots of calcification pixels. To ASN, they are also outliers for reconstruction even though they are very different with true μCs in shape, which are shown with orange circles in the right patch in Fig.3. While they are not hard to distinguish for discriminative models. Therefore, we cascade a deep classification network to further reduce the false positives.

We use ResNet50 [10] in FPR stage. Given an image, we use a simple threshold on the reconstruction residual map generated by ASN. For each connected component, we use the center as predicted location and the summed score of the reconstruction residual value as its ASN score. For each ASN prediction, a patch with size of 56×56 is cropped and resized to 224×224, and then fed into ResNet50. We use the product of both ASN and FPR scores as the final score.

4. Experiments

4.1. Implementation Details

Mammogram image is commonly stored as 12-bit or 14-bit data in DICOM format. To convert it into 8-bit gray image, we simply map all the raw pixels into $0 \sim 255$ linearly. For pre-processing, we first normalize the image to have the same pixel spacing of 70 μm. And then, we segment the breast region with Otsus method [21] and remove the background of the mammogram.

Table 2. Evaluations on in-house dataset (%).

Method	R@1	R@5	R@10	R@15	R@20
FPN FRCN	78.27	81.33	81.33	81.33	81.33
U-Net w FPR	84.90	88.06	88.67	88.67	88.67
Proposed	**85.31**	**90.71**	**92.24**	**92.65**	**92.76**

We implement proposed model with pytorch. ASN is trained from scratch with weights initialized by [9]. We use Adam [15] with a weight decay of 10^{-4} and a starting learning rate 0.001. The running averages of gradient and its square are 0.9 and 0.999 respectively. The margin parameter β in Eq.6 is set to 0.8, while the weighting parameters λ_p and λ_n are set to 1 and 0.1 respectively. During the training process, mammogram images are cropped into patches with size of 112×112 and fed into ASN. We do not use the whole images since they are usually of high resolution (e.g. $\sim 3500 \times 2500$ pixels), which is too large for memory limitation. We sample the positive and negative patches to be 1:1 to extract more proposals.

FPR model is pretrained by ImageNet. We use SGD with learning rate starting from 0.001. All predictions by ASN and all ground truth μCs are used to train FPR model without any extra sampling.

4.2. Datasets

We evaluate the performances on both a public dataset named INBreast [19] and an in-house dataset. There are several public mammogram datasets, i.e. MIAS [29], DDSM [11], INBreast and so on. We choose INBreast because the image quality and μCs annotations are relatively better. INBreast contains 115 cases with 410 mammogram images, in which 6880 individual calcifications have been found by two radiologists. After filtering the calcifications larger than 1mm, we pick up 5782 μCs for experiments. We randomly divide the dataset into training, validation and testing sets by 3:1:1. The detailed division is shown in supplementary materials.

We also collected an in-house dataset for further evaluation, which contains 439 cases and 1799 images. Images of different study years but from the same woman are taken as the same case. Two radiologists with experiences more than 10 years annotated the dataset. We pick up 7588 μCs identified by both radiologists as ground truth. We select 339 cases with 1386 images and 5479 μCs as training set, 50 cases with 208 images and 1129 μCs as validation set, 50 cases with 205 images and 980 μCs as testing set.

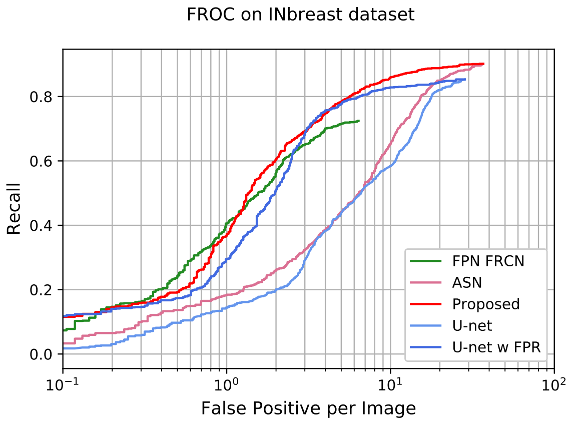


Figure 4. FROCs for INBreast dataset.

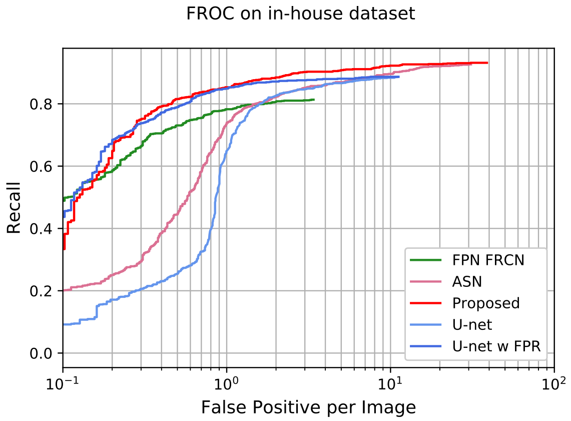


Figure 5. FROCs for in-house dataset.

4.3. Baseline Methods

For both datasets, we build two baselines:

FPN FRCN: Faster RCNN [23] with Feature Pyramid Network (FPN) [18]. FPN is a state-of-the-art detection model especially for small objects. We use ResNet50 as backbone. For each predicted bounding box, the center point is used for final evaluation.

U-Net w FPR: U-Net with FPR. U-Net [24] is proved to be effective for medical imaging segmentation. The skip connections are helpful for small object segmentation. Here we compare U-Net with ASN to verify the effectiveness of the proposed generative model. To deal with the extreme unbalanced samples, and also to be fair comparison with the proposed model, we design a two stage segmentation model similar with ASN. We first train a segmentation task using the same network structure with ASN supervised by cross-entropy loss. We also sample the positive and negative patches to be 1:1 to extract more proposals. We select the connected regions of the predicted mask as proposal similar to the proposed method. Then an FPR model is cascaded to reduce the false positives.

Table 3. Proposal evaluations on INBreast dataset (%).

Method	R@5	R@10	R@15	R@20	R@30
U-Net	43.39	58.35	72.20	82.02	85.32
ASN	**44.13**	**65.14**	**78.72**	**84.95**	**88.35**

Table 4. Proposal evaluations on in-house dataset (%).

Method	R@5	R@10	R@15	R@20	R@30
U-Net	86.63	88.37	88.67	88.67	88.67
ASN	**88.16**	**90.00**	**91.33**	**91.84**	**92.24**

4.4. Performances

We report the recalls at k false positive per image (simplified as R@k), where $k \in \{1, 5, 10, 15, 20\}$ for final models and $k \in \{5, 10, 15, 20, 30\}$ for proposal models. A μC is considered as recalled if there is at least one prediction point within 1mm of it.

As shown in Tab. 1 and 2, the proposed models outperform the state-of-the-art methods on both datasets. The FPN models suffer from relatively lower recalls. The main reason is that some μCs are extremely tiny ($\leq$ 5 pixels). The resolution of the finest prediction level of FPN is only $1/4$ with respect to the origin image. For the first 3 examples in Fig.6, FPN fails to detect either of them. Moreover, small size also means less positive anchors in RPN, which can lead to low recall. U-Net models can deal with the small μC size, since they predict pixelwisely. However, some obscure samples are still challenging and missed in the first stage, while the proposed models suffer less from them. Tab. 3 and 4 show the proposal quantitative evaluation results. ASN outperforms U-Net by around 3% on both datasets. According to Fig.4 and 5, both recall and false positive rate of ASN are higher than U-Net, which indicate that the generative method is more sensitive to μCs and other noisy outliers.

The fourth row of Fig.6 shows a vessel calcification example, which is predicted to be μCs by ASN. The vessel calcification regions can be seen as combinations of very local calcification pixels. However, they are very different to μCs in global patterns, which is not hard for FPR model to learn. In a nutshell, the generative model and the discriminative model is complementary in a way. Therefore, proposed model can take both advantages and achieve higher recall and lower false positive.

4.5. Ablation Study

To verify the effectiveness of t-test loss, we first train a plain reconstruction model with loss function that minimize the mean squared error between original image and reconstructed image. However, the model seems to collapse into a simple bluring function, where only a few high frequency contents appear in the reconstruction residual. It fails to identify most μCs. This phenomenon also indicates that

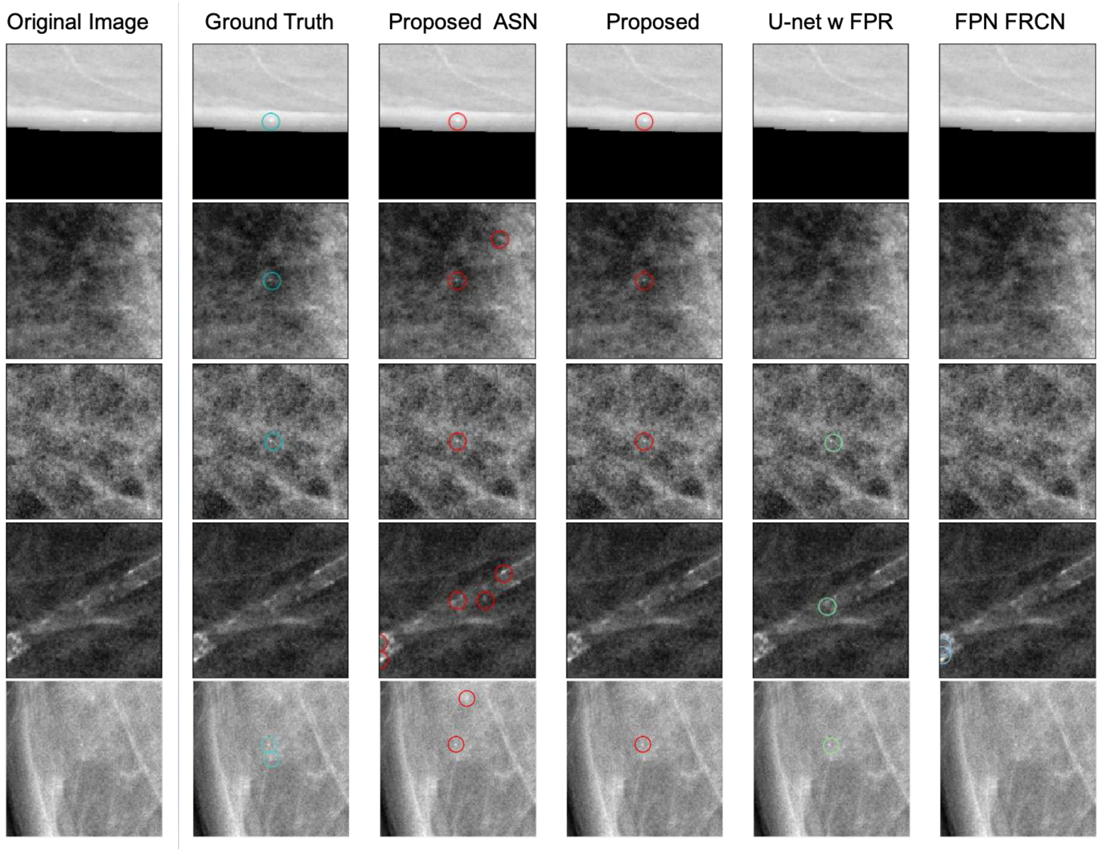


Figure 6. Comparison examples of detection results in INBreast dataset. The annotations by doctors are drawn in the second column with cyan circles. The third to sixth columns show the comparison between the proposed models and state-of-the-art models. Each detected calcification is shown by a circle in center of the predicted position.

proposed t-test loss is essential.

Then, to validate the necessity of regularizations, we set $\lambda_p = \lambda_n = 0$ and the loss function turns into Eq. 10, i.e. without the regularizations of variances of negative and positive residuals. The improvement of ours over the last line in as shown in Table 5 may be contributed to that the regularizations can avoid the estimations to be so scattered that the residuals are unstable.

In addition, we compare some variants to reveal the contribution of different components. In Eq.1, L1 distance is used to compute reconstruction residual. We replace it with L2 and SmoothL1 [23] to further study the influences. As shown in Table 5, such variants yield comparable results.

Table 5. Ablation study on INBreast dataset (%).

Method	R@1	R@5	R@10	R@15	R@20
L1	36.70	78.35	**85.96**	**88.26**	88.90
L2	**37.41**	**80.00**	85.41	87.98	88.53
SmoothL1	35.83	78.26	85.23	87.98	**89.08**
L1 w/o λ	29.90	72.02	83.30	86.06	86.88

5. Discussions and Conclusions

In this paper, we propose a novel model by cascading a discriminative model to a generative model to tackle the μC detection problem in mammogram images. The μCs are very tiny and also rare to the normal tissues, which is challenging for discriminative models. We first propose a novel generative model named Anomaly Separation Network (ASN) to extract proposals, and then train a classification network as False Positive Reduction (FPR) model. In ASN, a deep convolutional encoder-decoder network is applied to learn the reconstruction and a t-test loss function is proposed to train this network in a supervised way. Experiments on both public and in-house datasets demonstrate that our model outperforms previous state-of-the-art methods. However, it is still challenging for the proposed method when μCs are too close (the last row of Fig.6). In the future, we will work on it. Additionally, we will try to make the total pipeline training in an end-to-end manner.

Acknowledgments

This work was supported in part by NSFC-61625201, 61527804.

References

[1] Benign calcification detection in mammogram images.

[2] Ulrich Bick. *Mammography: How to Interpret Microcalcifications*. Springer Milan, 2014.

[3] A. Bria, N. Karssemeijer, and F. Tortorella. Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications. *Medical Image Analysis*, 18:241–252, 2014.

[4] Guanxiong Cai, Yanhui Guo, Yaqin Zhang, Genggeng Qin, Yuanpin Zhou, and Yao Lu. A fully automatic microcalcification detection approach based on deep convolution neural network. In *Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis*, 2018.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009.

[6] B Efron and T Hastie. Computer age statistical inference: Algorithms. *Evidence and Data Science, Institute of Mathematical Statistics Monographs*, 2016.

[7] G Enderlein. Hawkins, d. m.: Identification of outliers. chapman and hall, london new york 1980, 188 s., 14, 50. *Biometrical Journal*, 29(2):198–198, 1987.

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *computer vision and pattern recognition*, pages 770–778, 2016.

[11] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. *The Digital Database for Screening Mammography*. Springer Netherlands, 2001.

[12] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[13] Justin Johnson, Alexandre Alahi, and Fei Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016.

[14] Aya F Khalaf and Inas A Yassine. Novel features for microcalcification detection in digital mammogram images based on wavelet and statistical analysis. In *IEEE International Conference on Image Processing*, pages 1825–1829, 2015.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–13, 2015.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2014.

[17] LaurieDavies and UrsulaGather. The identification of multiple outliers. *Publications of the American Statistical Association*, 88(423):782–792, 1993.

[18] Tsung Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017.

[19] IC1 Moreira, I Amaral, I Domingues, A Cardoso, MJ Cardoso, and JS Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.

[20] Marimuthu Muthuvel, Balakumaran Thangaraju, and Gowrishankar Chinnasamy. Microcalcification cluster detection using multiscale products based hessian matrix via the tsallis thresholding scheme. *Pattern Recognition Letters*, 94:127–133, 2017.

[21] N Otsu. A threshold selection method from gray-level histogram. *IEEE Trans Smc*, 9(1):62–66, 1979.

[22] Raviv Raich and Alfred O Hero. Sparse image reconstruction for partially known blur functions. pages 637–640, 2006.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems*, pages 91–99, 2015.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[25] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

[26] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[27] Edward A Sickles. Breast cancer screening outcomes in women ages 40-49: clinical experience with service screening using modern mammography. *Journal of the National Cancer Institute. Monographs*, 22:90–104, 1996.

[28] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA: a cancer journal for clinicians*, 67:7–30, 2017.

[29] J. Suckling, J. Parker, and D. R. Dance. Themammographic image analysis society digital mammogram database. In *Int Work on Dig Mammography*, 1994.

[30] Volodymyr Turchenko, Eric Chalmers, and Artur Luczak. A deep convolutional auto-encoder with pooling - unpooling layers in caffe. *arxiv:1701.04949*, 2017.

[31] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 511, 2001.

[32] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[33] John Wright, Allen Y Yang, Arvind Ganesh, Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.