CROSS-VIEW ATTENTION NETWORK FOR BREAST CANCER SCREENING FROM MULTI-VIEW MAMMOGRAMS

Xuran Zhao ,Luyang Yu, Xun Wang

School of Computer and Information Engineering
Zhejiang Gongshang University, Hangzhou, China

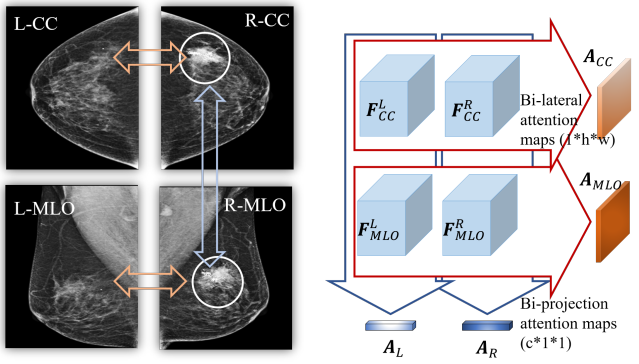
ABSTRACT

In this paper, we address the problem of breast caner detection from multi-view mammograms. We present a novel cross-view attention module (CvAM) which implicitly learns to focus on the cancer-related local abnormal regions and highlighting salient features by exploring cross-view information among four views of a screening mammography exam, e.g. asymmetries between left and right breasts and lesion correspondence between two views of the same breast. More specifically, the proposed CvAM calculates spatial attention maps based on the same view of different breasts to enhance bilateral asymmetric regions, and channel attention maps based on two different views of the same breast to enhance the feature channels corresponding to the same lesion in a single breast. CvAMs can be easily integrated into standard convolutional neural networks (CNN) architectures such as ResNet to form a multi-view classification model. Experiments are conducted on DDSM dataset, and results show that CvAMs can not only provide better classification accuracy over non-attention and single-view attention models, but also demonstrate better abnormality localization power using CNN visualization tools.

Index Terms— Mammogram Classification, Multi-view Learning, Attention Mechanism, Deep Learning

1. INTRODUCTION

Attention is one of the most influential ideas in deep learning. In computer vision tasks, the main objective of attention mechanism is to enable the model to focus on task-relevant local regions and feature channels of the input instead of treating all location and features as equally important [1, 2, 3]. In the area of medical image analysis, the importance of attention is even more intuitive: diagnosis are generally made by focused observation on small localized abnormal regions while most of the normal image part is less important. Recent researches in medical image analysis, such in skin lesion classification [4], thorax disease classification [5], and fetal ultrasound screening [6] have shown that incorporating attention modules in deep learning architectures can effectively improve classification accuracy and the ability to locate abnormalities compared to their standard non-attention convolutional neural networks (CNN) model counterparts. In these approaches, the attention maps were inferred from single images and cannot utilize cross-view information if multiple relevant inputs are provided.



(a) A typical 4-view mammography exam. Bi-lateral asymmetries and bi-projection co-occurrence of similar lesion draws clinicians attention.

(b) The proposed CvAM computes bi-lateral and bi-projection attention maps to mimic the clinical attention mechanism

Fig. 1: An illustration of attention mechanism in clinical mammogram interpretation and the bi-lateral and bi-projection attention in the proposed CvAM module

Breast cancer is the most common cancer and the second leading cancer-related cause of death among women in the world [7], and screening mammography is the main imaging method to detect breast cancer in early stage. A typical screening mammography exam is comprised of four images: for each breast, two images are taken from two different projection angles, namely Cranial-Caudal (CC) from above and mediolateral-oblique (MLO) taken horizontally. The 4 images are referred to as L-CC, L-MLO, R-CC and R-MLO respectively, as shown in Figure 1 (a). In clinical practices, cross-view information is important for finding abnormalities and making diagnosis. Where and what to focus in one mammogram image depend not only on the image itself, but also on other images. We refer this behavior as cross-view attention. We conclude two types of cross-view attention: first, asymmetric dense regions in the same projection of two different breasts attract attention, which we call bi-lateral attention; second, in the CC and MLO projections of the same breast, if a radiologist identifies a lesion in one projection, he/she will look for the same lesion in the other projection, and judge its malignancy based on the two observations, which we referred to as bi-projection attention.

Based on discussions above, we present in this paper a novel cross-view attention module (CvAM) to mimic the attention mechanism in clinical practices by utilizing information between different views to force a CNN model to focus on the cancer-related local

This work is supported in part by the National Natural Science Foundation of China under grand No.61702453, Natural Science Foundation of Zhejiang Province under grand No. LQ17F030001, Qianjiang Talent Program of Zhejiang Province under grand No. QJD1602021. The first author Xuran Zhao also works as artificial intelligence expert in Shenzhen XpectVision Technology Co. Ltd. The work is also supported by the corporational resources.

abnormal regions. As shown in Figure 1 (b), a CvAM takes the intermediate feature maps of 4 mammographic views as input and calculates spatial attention maps to identify asymmetric regions (as shown by red arrows) based on the same views of different breasts , and bi-projection attention maps to identify important feature channels (as shown by blue arrows) based on two different views of the same breast. Finally, each of the 4 input feature maps is multiplied by its corresponding spatial and channel attention maps to output a enhanced feature map. CvAMs can be easily integrated into standard CNN models such as ResNets [8] by inserting the them between consecutive convolutional stages. To demonstrate the effective of our approach, experiments of multi-view mammogram classification are conducted on the public DDSM [9] dataset. We show that improved accuracy and better interpretability are possible at the same time by using CvAM compared with non-attention baseline and and single-view attention models. We conjecture that the performance gain comes from accurate attention and noise reduction by effectively exploiting multi-view information.

In terms of related works, [10] and [11] addressed the problem of whole mammogram classification by combining full image and local lesion information. However, these approaches require pixel-level annotations from expert clinicians for model training, which are extremely expensive to acquire. Zhu et al. [12] formulated the mammogram classification as a multi-instance learning problem to force the malignancy prediction of a image dependent on sparse regions without pixel annotations. This method can be regarded as an implicit formulation of attention mechanism and does not require lesion annotation. Their approach is however based on single images. Multi-view mammogram classification was investigated in [10, 13, 14, 15], but these works typically concatenated multi-view features at different convolutional stages for fusion and did not involve attention mechanism. Woo et al. [3] proposed convolutional bottleneck attention module (CBAM), which use single-view information to infer channel and spatial attention maps. Our approach can be regarded as a multi-view extension of CBAM in the specific problem of mammogram classification.

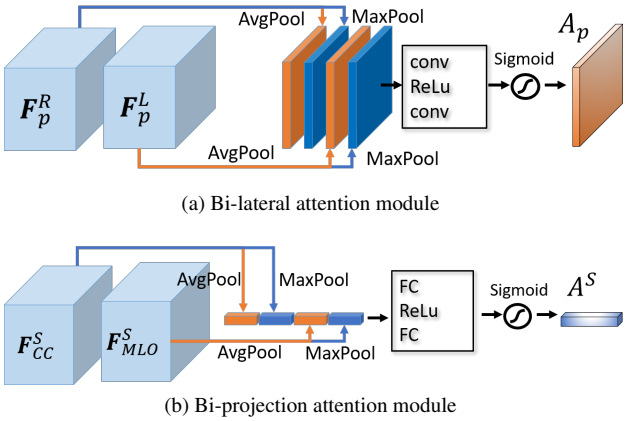
Our main contribution is two-folded: First, we propose a novel cross-view attention module (CvAM) specially designed for screening mammography classification to simulating the attention mechanism when clinicians interpret 4-view mammography exams. Second, we experimentally demonstrate that the proposed CvAM can boost the classification performance over the non-attention baseline and single-view attention models, and also improve lesion localization ability although even if no location annotation was provided in the training phase.

2. METHODOLOGY

In this section, we first introduce the proposed CvAM module and then describe a multi-view attention-based mammogram classification architecture by integrating CvAMs into a standard single-view, non-attention CNN classification model.

2.1. Cross-view Attention Module (CvAM)

The input of a CvAM includes intermediate feature maps $\mathbf{F}^L_{CC}$, $\mathbf{F}^L_{MLO}, \mathbf{F}^R_{CC}, \mathbf{F}^R_{MLO} \in \mathbb{R}^{c\times h\times w}$ for L-CC, L-MLO, R-CC, R-MLO views, respectively. A bi-lateral attention module calculates 2-D spatial attention maps $\{\mathbf{A}^{CC}, \mathbf{A}^{MLO}\} \in \mathbb{R}^{1\times h\times w}$ for CC and MLO views respectively based on the left and right feature maps of same projection, and a subsequent bi-projection attention module calculates 1-D channel attention maps $\{\mathbf{A}_L, \mathbf{A}_R\} \in \mathbb{R}^{c\times 1\times 1}$



(a) Bi-lateral attention module

(b) Bi-projection attention module

Fig. 2: An illustration of the bi-lateral and bi-projection attention sub-modules in CvAM.

for left and right breasts based on the two projection views of the same breast. The following describes the bi-lateral and bi-projection attention modules in detail.

Bi-lateral attention module. Given that the left and right breast images shoot from the same projection direction are generally symmetric, their feature maps are roughly spatially aligned (given one view is horizontally flipped). As a result, the bi-lateral attention module spatially combines left and right feature maps of the same projection to produce a spatial attention map to enhance the importance of asymmetric regions. Figure 2 (a) illustrates the generation of spatial attention map $\mathbf{A}_p$ for projection $p \in \{CC, MLO\}$. We first apply average-pooling and max-pooling along the channel axis of feature maps $\mathbf{F}^L_{CC}$ and $\mathbf{F}^R_{CC}$ separately and concatenate them into a $4 \times h \times w$ feature descriptor. Applying pooling operations along the channel axis is able to highlight informative regions [16]. The descriptor then passes through a 3×3 convolution layer with 4 output channels, an ReLU layer and another 3×3 convolution layer with 1 output channel, and a sigmoid layer finally compresses the output into an $1 \times h \times w$ spatial attention map $\mathbf{A}_p$. In short, the bi-lateral attention map of projection p is computed as:

$$\mathbf{A_p} = \sigma(C([AP(\mathbf{F}^L_p); MP(\mathbf{F}^L_p); AP(\mathbf{F}^R_p); MP(\mathbf{F}^R_p)])) \quad (1)$$

where σ is the sigmoid function, C represents the convolutional block, and AP/MP represents the average and max pooling respectively.

Bi-projection attention module. This module aims to aggregate information in CC and MLO projections of the same breast to generate attention. Figure 2 (b) illustrates the computation of attention map $\mathbf{A}^s$ from $\mathbf{F}^s_{CC}$ and $\mathbf{F}^s_{MLO}$, where $s \in \{L, R\}$. As it is difficult to find spatial correspondence between images acquired from two projection angles, we squeeze the spatial dimension of the input feature maps. Spatial average-pooling and max-pooling are applied to $\mathbf{F}^s_{CC}$ and $\mathbf{F}^s_{MLO}$ and the results are concatenated into a $4c \times 1 \times 1$ vector. This vector is then forwarded into a multi-layer perceptron (MLP) with one hidden layer to form the $c \times 1 \times 1$ channel attention map $\mathbf{A}^s$. In short, the channel attention map of side s is computed as:

$$\mathbf{A^s} = \sigma(MLP([AP(\mathbf{F}^s_{CC}); MP(\mathbf{F}^s_{CC}); AP(\mathbf{F}^s_{MLO}); MP(\mathbf{F}^s_{MLO})])), \quad (2)$$

where σ denotes the sigmoid function, and the MLP weights are shared between left and right breast.

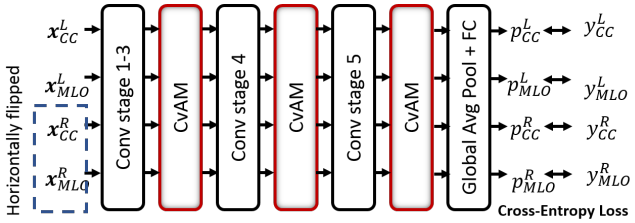


Fig. 3: General model architecture over multi-view mammogram classification

Finally, the feature map of each view are refined by element-wise multiplication with its corresponding spatial and channel attention maps plus 1:

$$\mathbf{F}_s^{p*} = (1 + \mathbf{A}_p) \otimes (1 + \mathbf{A}^s) \otimes \mathbf{F}_p^s \tag{3}$$

where $\otimes$ denotes element-wise multiplication, and $s \in \{L, R\}$ and $p \in \{CC, MLO\}$ denote a specific side and projection respectively. During multiplication, the attention values are broadcasted: channel attention values are copied along the spatial dimension, and vice versa. The whole process is summerized in Algorithm 1.

2.2. Multi-view mammogram classification with CvAM

A single-view non-attention CNN classification model such as ResNet [8] can be converted into a multi-view attention-based mammogram classification model by inserting CvAMs between consecutive convolutional stages. In this paper, we integrate three CvAMs into a standard ResNet-50 model, one CvAm after the 3^{rd}, 4^{th} and 5^{th} convolutional stage respectively, as shown in Fig. 3. A training sample includes 4 mammogram images and their corresponding labels, while the R-CC and R-MLO images are horizontally flipped to keep the same spatial layout as their left-breast counterparts. Each view is represented by $(\mathbf{x}_p^s, \mathbf{y}_p^s)_{s\in\{L,R\},p\in\{CC,MLO\}}$ where $\mathbf{x}_p^s$ is the mammogram image of side s and projection p, and $y_p^s \in \{0, 1\}$ is the ground truth malignancy. Note that the ground truth label is assigned to each breast, i.e., for each breast $y_{CC}^s = y_{MLO}^s$, but two breasts of the same patient can have different label. The 4 input images make independent forward pass through shared convolutional blocks in original ResNets, then jointly refined by CvAM. The 4 feature maps output by the last CvAM go through a global average pooling layer to squeeze the spatial dimension to 1×1, and a shared fully-connected layer is applied to each feature vector to make a prediction. Finally, cross-entropy losses are calculated between the 4 predictions and their corresponding label, and the losses are back-propagated to compute gradient.

Algorithm 1 CvAM

Input: Intermediate feature maps of 4 views in a screening mammography exam, $\{\mathbf{F}_L^{CC}, \mathbf{F}_L^{MLO}, \mathbf{F}_R^{CC}, \mathbf{F}_R^{MLO}\} \in \mathbb{R}^{c\times h\times w}$.
for $p = CC, MLO$ **do**
  Calculate bi-lateral attention $\mathbf{A}_p \in \mathbb{R}^{1\times h\times w}$ for projection p based on $\mathbf{F}_p^L$ and $\mathbf{F}_p^R$ according to Equation 1.
end for
for $s = L, R$ **do**
  Calculate bi-projection attention $\mathbf{A}^s \in \mathbb{R}^{c\times 1\times 1}$ for breast s based on $\mathbf{F}_{CC}^s$ and $\mathbf{F}_{MLO}^s$ according to Equation 2.
end for
for $p = CC, MLO$ **do**
  for $s = L, R$ **do**
    Calculate the refined feature map $\mathbf{F}_p^{s*}$ for side s and projection p according to Equation 3.
  end for
end for
Return: Refined feature maps $\mathbf{F}_L^{CC*}, \mathbf{F}_L^{MLO*}, \mathbf{F}_R^{CC*}, \mathbf{F}_R^{MLO*}$.

3. EXPERIMENTS

3.1. Dataset

The effectiveness of the proposed CvAM and multi-view attention-based mammogram classification network is evaluated with the public DDSM dataset [9], which contains 2620 exams of digitized film-screen mammography in 4 views. Other publicly available datasets such as INbreast [17] and MIAS [18] are either small in size or do not contain all 4 views of a mammographic exam. The exams in DDSM dataset are categorized into four classes, namely malignant (914 cases), benign (870 cases), benign-without-callback (141 cases) and normal (695 cases). We build a binary classifier to distinguish malignant (positive) and non-malignant (negative) breasts. Note that for most of the malignant cases, cancer only occurs in one breast, we label the other breast as negative in these cases. The dataset is divided into into train, validation and test sets according to a $0.8 : 0.1 : 0.1$ ratio. All the images are resized to a lower resolution of 416×416.

3.2. Compared methods

To show the benefits of the proposed multi-view attention mechanism, we compare the following single-view and multi-view approaches, involving or not attention mechanism:

- **ResNet50**. A standard ResNet50 model is used as a single-view, non-attention baseline.
- **ResNet50+feature concat**. In[10], different approaches were proposed and evaluated for merging CNN features of CC and MLO views to make a single decision. We adopt their best performing approach, i.e, the CC and MLO features output by the global average pooling layer are concatenated and send into a fully-connected layer for predicting the malignancy of a breast. This approach serves as a multi-view, non-attention baseline.
- **Resnet50+CBAM**. Our approach can be regarded as a multi-view extention of the convolutional bottleneck attention module (CBAM) proposed in [3], which also learns channel and spatial attention maps but only from single input feature. We integrate three CBAMs into the ResNet50 in the same way described in Section 2.2 to form a single-view attention-based architecture.
- **ResNet50+CvAM.** The proposed multi-view-attention-based architecture by incorporating CvAM modules into ResNet50.

3.3. Evaluation protocol

For each compared approach, receiver operating characteristic curve (ROC) is plotted as the true-positive rate versus the false-positive rate at various thresholds. The area under the curve (AUC) is used to measure and compare performances. For single-view methods and the proposed multi-view attention approach, different malignancy

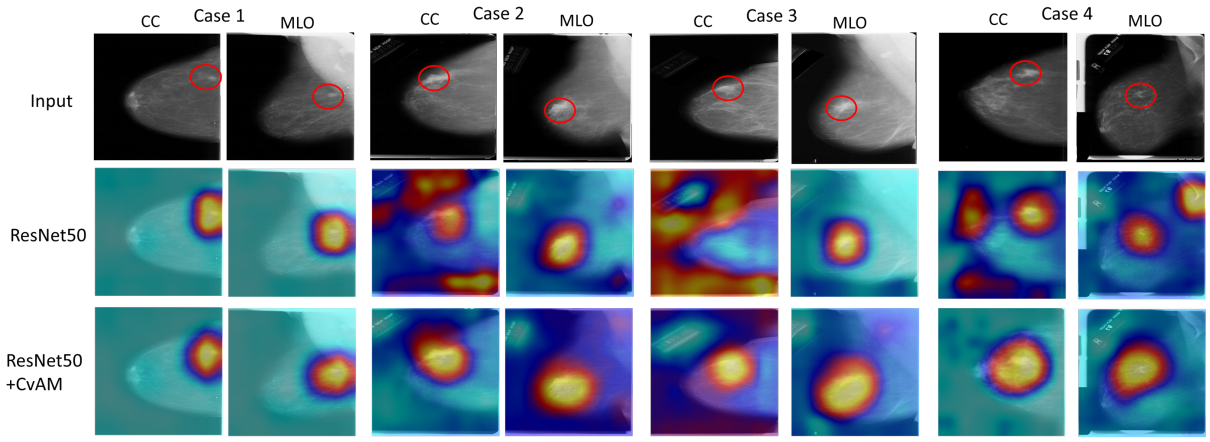


Fig. 4: CAM [19] visualization results of baseline ResNet50 (second row) and CvAM-embeded ResNet50 (third row). The first rows are input images and the lesions are marked by red circles.

Method	single image	CC-MLO fusion
ResNet50 Baseline	0.812	0.836
ResNet50+feature concat[10]	N.A.	0.830
ResNet50+CBAM[3]	0.825	0.845
ResNet50+CvAM (proposed)	**0.855**	**0.862**

Table 1: Comparison of AUCs of different approaches on DDSM test set. The AUCs are calculated in both single-image and CC-MLO score fusion mode.

scores are predicted for both CC and MLO views of a breast. Combining the prediction of CC and MLO views may increase performance because each view contain unique information. We used a simple approach of taking the average score of the two views, and report AUCs based on both single-image and score-fusion. For the multi-view feature fusion approach only the fusion score is provided, since the CC and MLO views have already been fused to generate a single score for a breast.

3.4. Results

Quantitative Analysis. The AUCs of four compared approaches were evaluated based on single image and CC-MLO score fusion, and is presented in Table 1. The ResNet50 baseline model achieved a reasonable AUC of 0.812 in single image mode, compared to the reported 0.77 AUC in [20] with the same dataset but using AlexNet [21] as classification model. A simple CC-MLO score fusion boosted the performance to 0.836, showing the importance of aggregating information between two views. The multi-view feature fusion approach obtained an AUC of 0.83, which is better than the single-image baseline but brings less improvement than the score-level fusion. By integrating attention mechanism, the single-view attention approach reaches AUCs of 0.825 and 0.845 in single-image and score fusion test respectively, showing moderate improvements over its non-attention baselines. Finally, the proposed multi-view attention model achieves the best single-image AUC of 0.853 of all compared approaches, which is 3% higher than the single-view attention model, on account of its ability to exploit both spatial and channel attention from multi-view training examples. When the CC and MLO predictions are fused, the AUC is further boosted to 0.862. This performance gain by score fusion is less significant as in single-view approaches, as the inter-view information is already used in the feature refinement process.

Network Visualization with CAM. Apart from high classification accuracy, a ideal mammogram classification model should be also capable of locating abnormalities even if no location annotations were provided in the training phase. To understand the influence of CvAM, we apply class activation mapping (CAM) [19] to the trained ResNet50 model and its CvAM-embedded counterpart to visualize the areas of the image which is most indicative of cancer. Figure 4 shows the CAM visualizations for CC and MLO views of 4 breasts in the test set which contain malignant lesions, marked out in red circles in the first row. In case 1, a tumor is clearly visible in a low-density breast. Both ResNet50 and the CvAM-embeded model successfully located the lesion. In more complicated breast images as case 2 and 3, the CAM of ResNet50 is able to located the lesion in one view but the activation in the other view widely spread out. The CvAM-embedded model located the lesion in both views, showing the ability of exploiting bi-projection information. In case 4, some high-density regions of normal tissues were activated in both views for ResNet50, while these false positives were suppressed in the CvAM-embedded activation map. We conjecture that the benefit comes from the bi-lateral attention module which references the information in the other breast.

4. CONCLUSION

We presents a novel cross-view attention module (CvAM) which implicitly learns to focus on the cancer-related local abnormal regions and highlighting salient features by exploring cross-view information among four views of a mammography exam. Experimental results show that the CvAMs-embedded multi-view attention model not only out-performs non-attention and single-view attention models in classification accuracy, but also provides better lesion locating ability. In future works, it is promising to extend the current work by: 1) incorporating multi-scale modeling to further improve whole mammogram classification; 2) apply mammogram registration approaches [22] to improve the spatial alignment of the left and right breast images , such that the proposed bi-lateral attention module may better learn the importance of asymmetric regions.

5. REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[2] Jianlong Fu, Heliang Zheng, and Tao Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.

[3] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[4] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen, "Attention residual learning for skin lesion classification," *IEEE transactions on medical imaging*, 2019.

[5] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[6] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[7] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal, "Cancer statistics, 2015," *CA: a cancer journal for clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W Philip Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th international workshop on digital mammography*. Medical Physics Publishing, 2000, pp. 212–218.

[10] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley, "Automated analysis of unregistered multi-view mammograms with deep learning," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2355–2365, 2017.

[11] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, 2019.

[12] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 603–611.

[13] Krzysztof J Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *arXiv preprint arXiv:1703.07047*, 2017.

[14] Lilei Sun, Junqian Wang, Zhijun Hu, Yong Xu, and Zhongwei Cui, "Multi-view convolutional neural networks for mammographic image classification," *IEEE Access*, vol. 7, pp. 126273–126282, 2019.

[15] Alan Joseph Bekker, Hayit Greenspan, and Jacob Goldberger, "A multi-view deep learning architecture for classification of breast microcalcifications," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 726–730.

[16] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[17] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[18] P SUCKLING J, "The mammographic image analysis society digital mammogram database," *Digital Mammo*, pp. 375–386, 1994.

[19] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[20] Sarah S Aboutalib, Aly A Mohamed, Wendie A Berg, Margarita L Zuley, Jules H Sumkin, and Shandong Wu, "Deep learning to distinguish recalled but benign mammography images in breast cancer screening," *Clinical Cancer Research*, vol. 24, no. 23, pp. 5902–5909, 2018.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[22] Saskia van Engeland, Peter Snoeren, JHCL Hendriks, and Nico Karssemeijer, "A comparison of methods for mammogram registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 11, pp. 1436–1444, 2003.