Mammo-Net: Integrating Gaze Supervision and Interactive Information in Multi-view Mammogram Classification

Changkai Ji[1,2], Changde Du[2], Qing Zhang[3], Sheng Wang[1,4,5], Chong Ma[6], Jiaming Xie[7], Yan Zhou[3], Huiguang He[1,2(✉)], and Dinggang Shen[1,5,8(✉)]

[1] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
{jichk,dgshen}@shanghaitech.edu.cn, huiguang.he@ia.ac.cn
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] Department of Radiology, Renji Hospital Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4] Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[5] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[6] School of Automation, Northwestern Polytechnical University, Xi'an, China
[7] Department of Computer Science, The University of Hong Kong, Hong Kong, China
[8] Shanghai Clinical Research and Trial Center, Shanghai, China

Abstract. Breast cancer diagnosis is a challenging task. Recently, the application of deep learning techniques to breast cancer diagnosis has become a popular trend. However, the effectiveness of deep neural networks is often limited by the lack of interpretability and the need for significant amount of manual annotations. To address these issues, we present a novel approach by leveraging both gaze data and multi-view data for mammogram classification. The gaze data of the radiologist serves as a low-cost and simple form of coarse annotation, which can provide rough localizations of lesions. We also develop a pyramid loss better fitting to the gaze-supervised process. Moreover, considering many studies overlooking interactive information relevant to diagnosis, we accordingly utilize transformer-based attention in our network to mutualize multi-view pathological information, and further employ a bidirectional fusion learning (BFL) to more effectively fuse multi-view information. Experimental results demonstrate that our proposed model significantly improves both mammogram classification performance and interpretability through incorporation of gaze data and cross-view interactive information.

Keywords: Mammogram classification · Gaze · Multi-view interaction · Bidirectional fusion learning

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_7.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 68–78, 2023.
https://doi.org/10.1007/978-3-031-43990-2_7

1 Introduction

Breast cancer is the most prevalent form of cancer among women and can have serious physical and mental health consequences if left unchecked [5]. Early detection through mammography is critical for early treatment and prevention [19]. Mammograms provide images of breast tissue, which are taken from two views: the cranio-caudal (CC) view, and the medio-lateral oblique (MLO) view [4]. By identifying breast cancer early, patients can receive targeted treatment before the disease progresses.

Deep neural networks have been widely adopted for breast cancer diagnosis to alleviate the workload of radiologists. However, these models often require a large number of manual annotations and lack interpretability, which can prevent their broader applications in breast cancer diagnosis. Radiologists typically focus on areas with breast lesions during mammogram reading [11,22], which provides valuable guidance. We propose using real-time eye tracking information from radiologists to optimize our model. By using gaze data to guide model training, we can improve model interpretability and performance [24].

Radiologists' eye movements can be automatically and unobtrusively recorded during the process of reading mammograms, providing a valuable source of data without the need for manual labeling. Previous studies have incorporated radiologists' eye-gaze as a form of weak supervision, which directs the network's attention to the regions with possible lesions [15,23]. Leveraging gaze from radiologists to aid in model training *not only* increases efficiency and minimizes the risk of errors linked to manual annotation, *but also* can be seamlessly implemented without affecting radiologists' normal clinical interpretation of mammograms.

Mammography primarily detects two types of breast lesions: masses and microcalcifications [16]. The determination of the benign or malignant nature of masses is largely dependent on the smoothness of their edges [13]. The gaze data can guide the model's attention towards the malignant masses. Microcalcifications are small calcium deposits which exhibit irregular boundaries on mammograms [9]. This feature makes them challenging to identify, often leading to missed or false detection by models. Radiologists need to magnify mammograms to differentiate between benign scattered calcifications and clustered calcifications, the latter of which are more likely to be malignant and necessitate further diagnosis. Leveraging gaze data can guide the model to locate malignant calcifications.

In this work, we propose a novel diagnostic model, namely Mammo-Net, which integrates radiologists' gaze data and interactive information between CC-view and MLO-view to enhance diagnostic performance. To the best of our knowledge, this is the first work to integrate gaze data into multi-view mammography classification. We utilize class activation map (CAM) [18] to calculate the attention maps for the model. Additionally, we apply pyramid loss to maintain consistency between radiologists' gaze heat maps and the model's attention maps at multiple scales of the pyramid [1]. Our model is designed for single-breast cases. Mammo-Net extracts multi-view features and utilizes transformer-based attention to mutualize information [21]. Furthermore, there are differences

between multi-view mammograms of the same patient, arising from variations in breast shape and density. Capturing these multi-view shared features can be a challenge for models. To address this issue, we develop a novel method called bidirectional fusion learning (BFL) to extract shared features from multi-view mammograms.

Our contributions can be summarized as follows:

- We emphasize the significance of low-cost gaze to provide weakly-supervised positioning and visual interpretability for the model. Additionally, we develop a pyramid loss that adapts to the supervised process.
- We propose a novel breast cancer diagnosis model, namely Mammo-Net. This model employs transformer-based attention to mutualize information and uses BFL to integrate task-related information to make accurate predictions.
- We demonstrate the effectiveness of our approach through experiments using mammography datasets, which show the superiority of Mammo-Net.

2 Proposed Method

2.1 Overall Architecture

The pipeline of Mammo-Net is illustrated in Fig. 1. Mammo-Net feeds two-view mammograms of the same breast into two ResNet-style [7] CNN branch networks. We use several ResNet blocks pre-trained on ImageNet [3] to process mammograms. Then, we use global average pooling (GAP) and fully connected layers to compute the feature vectors produced by the model. Before the final residual block, we employ cross-view attention to mutualize multi-view information. Our proposed method employs BFL to effectively fuse multi-view information to improve diagnostic accuracy. Additionally, by integrating gaze data from radiologists, our proposed model is able to generate more precise attention maps. The fusion network combines multi-view feature representations using a stack of linear-activation layers and a fully connected layer, resulting in a classification output.

2.2 Gaze Supervision

In this module, we utilize CAM to calculate the attention map for the network by examining gradient-based activations in back-propagation. After that, we employ pyramid loss to make the network attention being consistent with the supervision of radiologists' gaze heat maps, guiding the network to focus on the same lesion areas as the radiologists. This module guides the network to accurately extract pathological features.

Class Activation Map. At the final convolutional layer of our model, the activation of the ith feature map $f_i(x, y)$ at coordinates (x, y) is associated with a weight w_i^k for class k. This allows us to generate the attention map H^k for class k as:

$$H^k = \sum_i w_i^k f_i(x, y). \tag{1}$$

Pyramid Loss. To enhance the learning of important attention areas, we propose a pyramid loss constraint that requires consistency between the network and gaze attention maps. The pyramid loss is based on using a pyramid representation of the attention map:

$$\mathcal{L}_{Pyramid} = \sum_{l}^{L} ||(Z(G_l(H)))^+ - (Z(G_l(R)))^+||_2, \tag{2}$$

where H is the network attention map generated by the CAM and R is the radiologist's gaze heat map. $G_l(\cdot)$ represents the feature map at the lth level of the Gaussian pyramid, obtained by downsampling $G_{l-1}(\cdot)$ using a Gaussian kernel, where $G_1(R) = R$. Z means to perform Layernorm and ReLU activation on each feature map. This focuses the consistency loss on the more important pathological regions. The positive part of the normalized $Z(R)$, denoted as $Z(R)^+$, indicates the network focuses on the lesions where the radiologist spent most time reading. The minimization of the pyramid loss involves calculating the mean

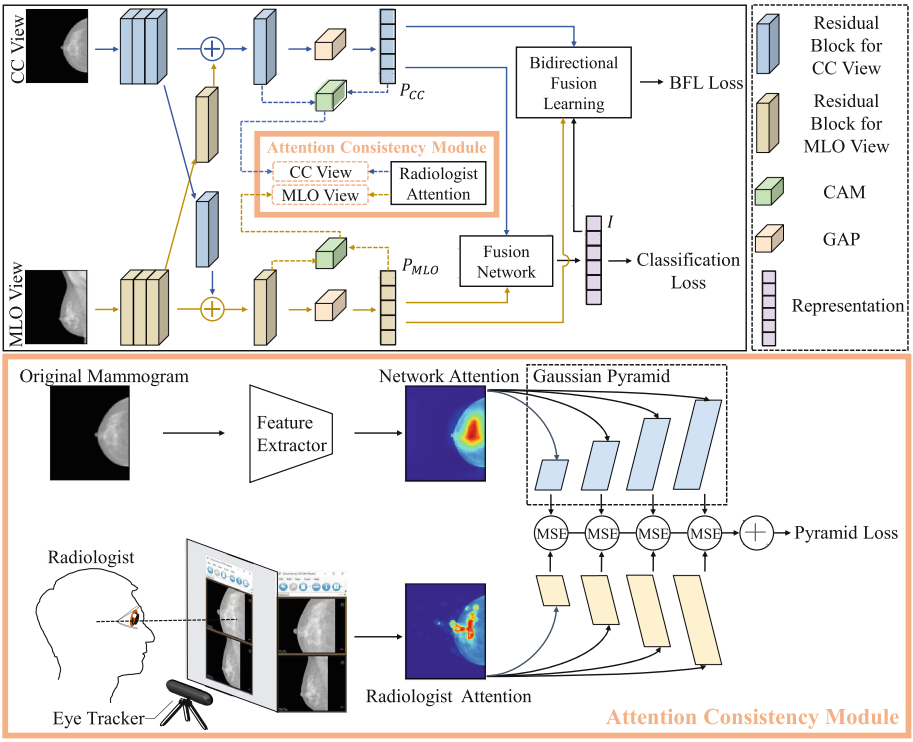

Fig. 1. Mammo-Net consists of two components: a multi-view classification network (upper half) and an attention consistency module (lower half). The classification network interacts multi-view information, while the attention consistency module provides positional supervision.

square error (MSE) between the attention maps generated by the radiologist and the model at each level of the Gaussian pyramid. This allows the model to mimic the attention of radiologists and enhance diagnostic performance.

Moreover, the pyramid representation enables the model to learn from the important pathological regions on which radiologists are focusing, without the need for precise pixel-level information. Layernorm is also employed to address the issue of imprecise gaze data. This reduces noise in the consistency process by performing consistency loss only in the regions where radiologist spent most time.

2.3 Interactive Information

Transformer-Based Mutualization Model. We use transformer-based attention to mutualize information from the two views at the level of the spatial feature map. For each attention head, we compute embeddings for the source and target pixels. Our model does not utilize positional encoding, as it encodes the relative position of each pixel and is not suitable for capturing information between different views of mammograms [21]. The target view feature maps are transformed into Q, the source view feature maps are transformed into K, and the original source feature maps are transformed into V. We can then obtain a weighted sum of the features from the source view for each target pixel using [21]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{3}$$

Subsequently, the output is transformed into attention-based feature maps X and mutualized with the feature maps Y from the other view. The mutualized feature maps are normalized and used for subsequent calculations:

$$Z = Norm(Y + Linear(X)). \tag{4}$$

Bidirectional Fusion Learning. To enable the fusion network to retain more of the shared features between the two views and filter out noise, we propose to use BFL to learn a fusion representation that maximizes the cross-view mutual information. The optimization target is to generate a fusion representation I from multi-view representations p_v, where $v \in \{cc, mlo\}$. We employ the Noise-Contrastive Estimation framework [6] to maximize the mutual information, which is a contrastive learning framework:

$$\mathcal{L}(I, \boldsymbol{P}_v) = -\mathbb{E}_{\boldsymbol{P}} \left[log \frac{s(I, p_v^i)}{\sum_{p_v^j \in \boldsymbol{P}_v} s(I, p_v^j)} \right], \tag{5}$$

where $s(I, p_v)$ evaluates the correlation between multi-view fused representations and single-view representations [17]:

$$\begin{aligned} s(I, p_v) &= exp\left(\overline{p_v}\left(\overline{N(I)}\right)^T\right), \\ \overline{p_v} = \frac{p_v}{||p_v||_2}, &\qquad \overline{N(I)} = \frac{N(I)}{||N(I)||_2}, \end{aligned} \tag{6}$$

where $N(I)$ is a reconstruction of p_v generated by a fully connected network N from I and the Euclidean norm $||\cdot||_2$ is applied to obtain unit-length vectors. In contrastive learning, we consider the same patient mammograms as positive samples and those from different patient mammograms in the same batch $\tilde{\boldsymbol{P}}_v^i = \boldsymbol{P}_v \backslash \{p_v^i\}$ as negative samples [17]. Minimizing the similarity between the same patient mammograms enables the model to learn shared features. Maximizing the dissimilarity between different patient mammograms enhances the model's robustness.

In short, we require the fusion representation I to reversely reconstruct multi-view representations p_v so that more view-invariant information can be passed to I. By aligning the prediction $N(I)$ to p_v, we enable the model to decide how much information it should receive from each view.

The overall loss function for this module is the sum of the losses defined for each view:

$$\mathcal{L}_{BFL} = \mathcal{L}_I^{cc} + \mathcal{L}_I^{mlo}. \tag{7}$$

2.4 Loss Function

We use binary cross entropy loss (BCE) between the network prediction and the ground-truth as the classification loss. In conclusion, we have proposed a total of three loss functions to guide the model training: $\mathcal{L}_{BCE}$, $\mathcal{L}_{BFL}$, and $\mathcal{L}_{Pyramid}$. The overall loss function is defined as the sum of these three loss functions, with coefficients λ and μ used to adjust their relative weights:

$$\mathcal{L}_{overall} = \mathcal{L}_{BCE} + \lambda \mathcal{L}_{Pyramid} + \mu \mathcal{L}_{BFL}. \tag{8}$$

3 Experiments and Results

3.1 Datasets

Mammogram Dataset. Our experiments were conducted on CBIS-DDSM [12] and INbreast [16]. The CBIS-DDSM dataset contains 1249 exams that have been divided based on the presence or absence of masses, which we used to perform mass classification. The INbreast dataset contains 115 exams with both masses and micro-calcifications, on which we performed benign and malignant classification. We split the INbreast dataset into training and testing sets in a 7:3 ratio. It is worth noting that the official INbreast dataset does not provide image-level labels, so we obtained these labels following Shen et al. [20].

Eye Gaze Dataset. Eye movement data was collected by reviewing all cases in INbreast using a Tobii Pro Nano eye tracker. The scenario is shown in Appendix and can be accessed at https://github.com/JamesQFreeman/MicEye. Participated radiologist has 11 years of experience in mammography screening.

3.2 Implementation Details

We trained our model using the Adam optimizer [10] with a learning rate of 10^{-4} (partly implemented by MindSpore). To overcome the problem of limited data, we employed various data augmentation techniques, including translation, rotation, and flipping. To address the problem of imbalanced classes, we utilized a weighted loss function that assigns higher weights to malign cases in order to balance the number of benign and malign cases. The coefficients λ and μ of $\mathcal{L}_{overall}$ were set to 0.5 and 0.2, respectively, based on 5-fold cross validation on the training set. The network was trained for 300 epochs. We used Accuracy (ACC) and the Area Under the ROC Curve (AUC) [25] as our evaluation metrics, and we selected the final model based on the best validation AUC. Considering the relatively small size of our dataset, we used ResNet-18 as the backbone of our network.

3.3 Results and Analysis

Table 1. Ablation study of key components of Mammo-Net, and comparison of different models in terms of AUC and ACC. "BFL" denotes "Bidirectional Fusion Learning", and "RA" denotes "Radiologist Attention".

Dataset	Model	AUC	ACC
CBIS-DDSM	Lopez et al. [14]	0.739	0.754
	Tulder et al. [2]	0.802	0.811
	Xian et al. [26]	0.812	0.735
	MLO-view	0.701	0.763
	CC-view	0.721	0.754
	Cross-view	0.809	0.838
	Cross-view+BFL	**0.821**	**0.864**
INbreast	Wang et al. [23]	0.806	0.756
	Jiang et al. [8]	0.819	0.793
	Lopez et al. [14]	0.793	0.830
	Xian et al. [26]	0.859	0.791
	MLO-view	0.663	0.716
	CC-view	0.650	0.704
	Cross-view	0.762	0.755
	Cross-view+BFL	0.786	0.812
	Cross-view+RA	0.864	0.830
	Cross-view+BFL+RA (Mammo-Net)	**0.889**	**0.849**

Performance Comparison. As shown in Table 1, we compare our model to other methods and find that our model performs better. Lopez et al. [14] proposed the use of hypercomplex networks to mimic radiologists. By leveraging the properties of hypercomplex algebra, the model is able to continually process two mammograms together. Lee et al. [26] proposed a 2-channel approach that utilizes a Gaussian model to capture the spatial correlation between lesions across two views, and an LT-GAN to achieve a robust mammography classification.

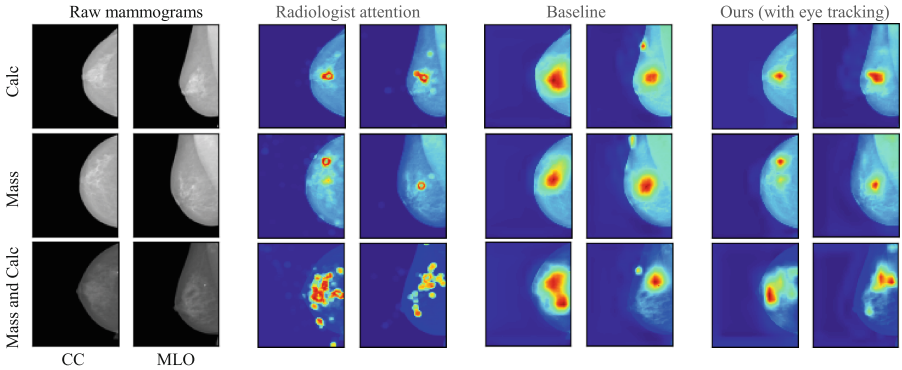

Fig. 2. Comparative visualization of mammography diagnosis with and without gaze supervision. After integrating gaze supervision, the model's capability in localizing lesions becomes more precise.

We also compare our model with other methods that use eye movement supervision as shown in Table 1. The GA-Net [23] proposed a ResNet-based model with class activation mapping guided by eye gaze data. We developed a multi-view model using this approach for a fair comparison, and found that our method performed better. We believe that one possible reason for the inferior performance of GA-Net compared to Mammo-Net might be the use of a simple MSE loss by GA-Net, which neglects the coarse nature of the gaze data. Jiang et al. [8] proposed a Double-model that fuses gaze maps with original images before training. However, this model did not consider the gap between research and clinical workflow. This model requires gaze input during both the training and inference stages, which limits its practical use in hospitals without eye-trackers. In contrast, our method does not rely on gaze input during inference stage.

Visualization. Figure 2 illustrates the visualization of our proposed model on three representative exams from the INbreast dataset that includes masses, calcifications, and a combination of both. For each exam, we present gaze heat maps generated from eye movement data. The preprocessing process is shown in Fig. 5 (see Appendix). To make an intuitive comparison, we exhibit attention maps generated by the model under both unsupervised and gaze-supervised

cases. Each exam is composed of two views, i.e., the CC-view and the MLO-view. More exams can be found in Fig. 6 (see Appendix).

The results of the visualization demonstrate that the model's capability in localizing lesions becomes more precise when radiologist attention is incorporated in the training stage. The pyramid loss improves the model's robustness even when the radiologist's gaze data is not entirely focused on the breast. This intuitively demonstrates the effectiveness of training the model with eye-tracking supervision.

Ablation Study. We perform an ablation analysis to assess each component (radiologist attention, cross-view attention and BFL) in Mammo-Net. Table 1 suggests that each part of the proposed framework contributes to the increased performance. This shows the benefits of adapting the model to mimic the radiologist's decision-making process.

4 Conclusion and Discussion

In this paper, we have developed a breast cancer diagnosis model to mimic the radiologist's decision-making process. To achieve this, we integrate gaze data as a form of weak supervision for both lesion positioning and interpretability of the model. We also utilize transformer-based attention to mutualize multi-view information and further develop BFL to fully fuse multi-view information. Our experimental results on mammography datasets demonstrate the superiority of our proposed model. In future work, we intend to explore the use of scanning path analysis as a means of obtaining insights into the pathology-relevant regions of lesions.

Acknowledgements. This work was supported in part by The Key R&D Program of Guangdong Province, China (grant number 2021B0101420006), National Natural Science Foundation of China (grant numbers 62131015, 82272072), Science and Technology Commission of Shanghai Municipality (STCSM) (grant number 21010502600), and the CAAI-Huawei MindSpore Open Fund.

References

1. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA Eng. **29**(6), 33–41 (1984)
2. Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X.: Depth estimation for colonoscopy images with self-supervised learning from videos. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 119–128. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_12
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)

4. Frazer, H.M., Qin, A.K., Pan, H., Brotchie, P.: Evaluation of deep learning-based artificial intelligence techniques for breast cancer detection on mammograms: results from a retrospective study using a breastscreen victoria dataset. J. Med. Imaging Radiat. Oncol. **65**(5), 529–537 (2021)
5. Giaquinto, A.N., Miller, K.D., Tossas, K.Y., Winn, R.A., Jemal, A., Siegel, R.L.: Cancer statistics for African American/black people 2022. CA Cancer J. Clin. **72**(3), 202–229 (2022)
6. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 297–304. JMLR Workshop and Conference Proceedings (2010)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Jiang, H., et al.: Eye tracking based deep learning analysis for the early detection of diabetic retinopathy: a pilot study. Available at SSRN 4247845 (2023)
9. Jørgensen, K.J., et al.: Breast-cancer screening-viewpoint of the iarc working group. New Engl. J. Med. **373**, 1478 (2015)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kundel, H.L., Nodine, C.F., Krupinski, E.A., Mello-Thoms, C.: Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. Acad. Radiol. **15**(7), 881–886 (2008)
12. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data **4**(1), 1–9 (2017)
13. Li, Z., et al.: Domain generalization for mammography detection via multi-style and multi-view contrastive learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 98–108. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_10
14. Lopez, E., Grassucci, E., Valleriani, M., Comminiello, D.: Multi-view breast cancer classification via hypercomplex neural networks. arXiv preprint arXiv:2204.05798 (2022)
15. Ma, C., et al.: Eye-gaze-guided vision transformer for rectifying shortcut learning. IEEE Trans. Med. Imaging (2023)
16. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012)
17. Oord, A.V.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Ouyang, X., et al.: Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. IEEE Trans. Med. Imaging **40**(10), 2698–2710 (2020)
19. Selvi, R.: Breast Diseases: Imaging and Clinical Management. Springer, Heidelberg (2014). https://doi.org/10.1007/978-81-322-2077-0
20. Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh, W.: Deep learning to improve breast cancer detection on screening mammography. Sci. Rep. **9**(1), 12495 (2019)
21. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)

22. Voisin, S., Pinto, F., Xu, S., Morin-Ducote, G., Hudson, K., Tourassi, G.D.: Investigating the association of eye gaze pattern and diagnostic error in mammography. In: Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment, vol. 8673, p. 867302. SPIE (2013)
23. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Trans. Med. Imaging **41**(7), 1688–1698 (2022)
24. Wu, C.C., Wolfe, J.M.: Eye movements in medical image perception: a selective review of past, present and future. Vision **3**(2), 32 (2019)
25. Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging **39**(4), 1184–1194 (2019)
26. Xian, J., Wang, Z., Cheng, K.-T., Yang, X.: Towards robust dual-view transformation via densifying sparse supervision for mammography lesion matching. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 355–365. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_34