IHRRB-DINO: Identifying High-Risk Regions of Breast Masses in Mammogram Images Using Data-Driven Instance Noise (DINO)

Mahmoud SalahEldin Kasem[1,4*], Abdelrahman Abdallah[2,3*], Ibrahim Abdelhalim[5⋆], Norah Saleh Alghamdi[6], Sohail Contractor[5], and Ayman El-Baz[5]

[1] Department of Multimedia, Assiut University, Egypt
[2] Department of Information Technology, Assiut University, Egypt
[3] Department of Computer Science, University of Innsbruck, Austria
[4] Information and Communication Engineering, Chungbuk University, South Korea
[5] University of Louisville, Louisville, Kentucky, USA
[6] Princess Nourah bint Abdulrahman University, Saudi Arabia

Abstract. In this paper, we introduce IHRRB-DINO, an advanced model designed to assist radiologists in effectively detecting breast masses in mammogram images. This tool is specifically engineered to highlight high-risk regions, enhancing the capability of radiologists in identifying breast masses for more accurate and efficient assessments. Our approach incorporates a novel technique that employs Data-Driven Instance Noise (DINO) for Object Localization, which significantly improves breast mass localization. This method is augmented by data augmentation using instance-level noise during the training phase, focusing on refining the model's proficiency in precisely localizing breast masses in mammographic images. Rigorous testing and validation conducted on the BI-RADS dataset using our model, especially with the Swin-L backbone, have demonstrated promising results. We achieved an Average Precision (AP) of 46.96, indicating a substantial improvement in the accuracy and consistency of breast cancer (BC) detection and localization. These results underscore the potential of IHRRB-DINO in contributing to the advancements in computer-aided diagnosis systems for breast cancer, marking a significant stride in the field of medical imaging technology.

Keywords: Breast Cancer Detection · Object Localization · Data-Driven Instance Noise (DINO) · Transformer.

1 Introduction

Breast cancer continues to pose one of the most significant challenges in public health, standing as the foremost cause of cancer-related deaths among women across the globe. Its prevalence, particularly in the United States, as the most

⋆ Equal contribution.

common cancer type among women, highlights an urgent need for effective early detection and intervention methods. The criticality of this issue is accentuated by the pivotal role of early detection in reducing mortality rates. Techniques such as mammography play a vital role in this regard, enabling the initiation of treatments at a stage where the cancer is most amenable to management and expanding the range of possible therapeutic options. This early intervention is key to enhancing patient outcomes and survival rates[9, 15]. The American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS) stratifies breast density into four precise categories, wherein elevated categories are associated with a heightened risk of breast cancer[5, 2]. Figure 1 illustrates samples of the dataset.

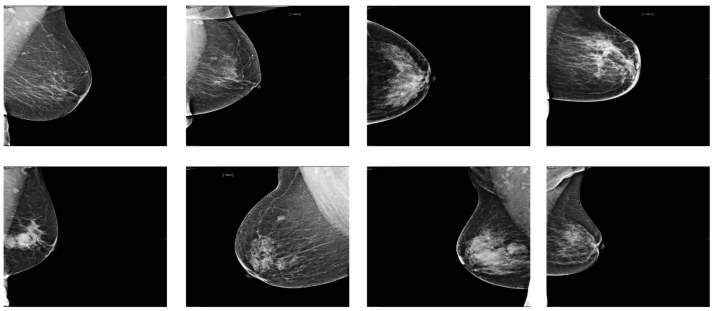

Fig. 1: Examples of the Dataset.

The application of deep learning (DL) models has markedly transformed medical diagnostics and screening, most notably in the detection and risk prediction of breast cancer [9, 15]. Recent advancements have leveraged computational methodologies to enhance the precision and efficacy of breast cancer screening protocols. Numerous studies [13, 12, 17, 19] have contributed to this domain by employing DL models for both the identification of present cancerous lesions and the projection of future risk [20, 1].

The advent of Vision Transformers (ViTs) [18] has notably revolutionized the field of medical imaging and diagnostics, challenging the erstwhile preeminence of Convolutional Neural Networks (CNNs) by adopting self-attention mechanisms for image analysis [4]. This technique enables an exhaustive evaluation of spatial relationships, which are pivotal in the accurate detection of breast cancer, thereby significantly advancing diagnostic accuracy and improving patient outcomes. ViTs herald a departure from the constraints associated with CNNs, augmenting the detection and diagnosis capabilities for breast cancer through superior pattern recognition competencies.

This progress has the potential to redefine the standards for breast cancer screening and diagnosis, offering a more sophisticated and advanced method for identifying cancers at their earliest stages. This research paper explores the

impact of Vision Transformers in breast cancer detection AND localization, comparing their effectiveness with traditional CNN-based methods. Through a comprehensive analysis of existing methodologies, this study aims to clear the future direction for using deep learning innovations to tackle one of the most urgent health challenges. The goal of this research is to contribute to worldwide efforts to lessen the impact of breast cancer and improve the success of early detection strategies.

2 Method

In our study, we developed an Innovative end-to-end Vision Transformer model specifically customized for the Detection of Breast Cancer. Our architecture comprises four key components: a backbone, a multi-layer Transformer encoder, a multi-layer Transformer decoder, and multiple prediction heads. This design allows for efficient and precise detection of breast cancer signatures in medical imaging, as shown in Figure 2

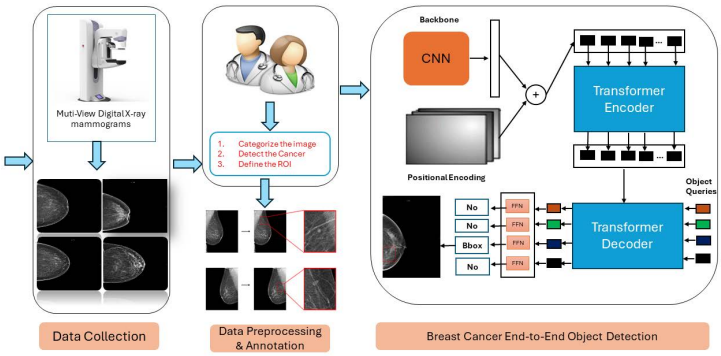

Fig. 2: Comprehensive Visualization of the Identifying High-Risk Regions of Breast Masses in Mammogram Images

2.1 Vision Tranmsformer

Vision Transformers (ViTs), as proposed in several key studies like [18, 4, 10], have significantly transformed the way visual data is processed. These models approach image analysis by breaking down images into a sequence of patches, which allows for the effective capture of long-range dependencies within the image. This is primarily achieved through the use of self-attention mechanisms, a concept further elaborated by P Shaw[16].

For a given input image, denoted as $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channel count respectively, the image is divided into N patches. Each patch is of dimensions $P \times P \times C$. These patches are then linearly embedded and processed through the attention mechanism of the transformer model. Vision Transformers employ a multi-head attention mechanism, enabling the model to simultaneously focus on different segments of the input data from multiple representation subspaces.

In terms of the mathematical formulation, given the embedded patch representations $\mathbf{e} \in \mathbb{R}^{N \times D}$, the attention mechanism for a single head is described by the following equation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

In this equation, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively. These are derived from the embedded patch representations. The term d_k denotes the dimensionality of the key. After computing the attention for each head, the outputs are then combined as follows:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_H)\mathbf{W}^O, \tag{2}$$

In this multi-head attention formulation, head_i is defined as $\text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$, where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ are the projection matrices for the i-th head. $\mathbf{W}^O$ is the output projection matrix. This structure allows Vision Transformers to effectively process visual data by focusing on different aspects of the input through various attention heads.

2.2 Model

The DINO model leverages a sophisticated architecture designed for breast cancer detection, incorporating a multi-layer Transformer[18] encoder and decoder alongside a backbone and multiple prediction heads. This setup enables the extraction of multi-scale features from images using advanced networks like ResNet[6] or Swin Transformer[10], which are then enhanced through a novel mixed query selection strategy. This strategy introduces deformable attention[24] to refine anchor boxes and classification outcomes. A crucial aspect of this model is the introduction of the Contrastive Denoising (CDN) training approach, enhancing model discrimination by handling hard negative samples.

The CDN approach is pivotal, defined by two hyper-parameters, λ_1 and λ_2, which control the noise scale for positive and negative queries, respectively. Positive queries, within a noise scale smaller than λ_1, aim to reconstruct corresponding ground truth boxes, while negative queries, within a noise scale between λ_1 and λ_2, predict "no object". This bifurcation is critical for improving the model's accuracy by teaching it to reject irrelevant anchors, detailed through the equation:

$$\mathrm{ATD}(k) = \frac{1}{k} \sum_{\mathrm{topK}} \{\|b_0 - a_0\|_1, \|b_1 - a_1\|_1, ..., \|b_{N-1} - a_{N-1}\|_1\},$$

where $\|b_i - a_i\|_1$ represents the L1 distance between the ground truth box b_i and its corresponding anchor a_i, emphasizing the model's efficiency in selecting quality anchors by minimizing confusion and duplicate predictions.

We also use *look forward twice* method innovates by allowing the parameters of a given layer (layer i) to be updated not only based on the losses of that layer but also incorporating the losses from the subsequent layer (layer $i+1$). This bidirectional flow of gradient information ensures that the refinement of a predicted bounding box (b^i_{pred}) is informed by both the quality of the initial bounding box (b_{i-1}) and the predicted box offset (Δb_i), thereby enhancing the precision of the model's predictions.

Given an input box b_{i-1} for the i-th layer, we obtain the final prediction box b^i_{pred} by the following process:

$$\Delta b_i = \mathrm{Layer}_i(b_{i-1}), \qquad b'_i = \mathrm{Update}(b_{i-1}, \Delta b_i), \tag{3}$$

$$b_i = \mathrm{Detach}(b'_i), \qquad b^i_{\mathrm{pred}} = \mathrm{Update}(b'_{i-1}, \Delta b_i), \tag{4}$$

where b'_i is the undetached version of b_i. The term $\mathrm{Update}(\cdot, \cdot)$ is a function that refines the box b_{i-1} by the predicted box offset Δb_i. This approach, termed as *look forward twice*, allows for the parameters of layer i to be influenced by the losses of both layer i and layer $i+1$.

The architecture and operational flow of our proposed model are illustrated in Figure 3, showcasing the integration of advanced deep learning techniques for enhanced breast cancer detection.

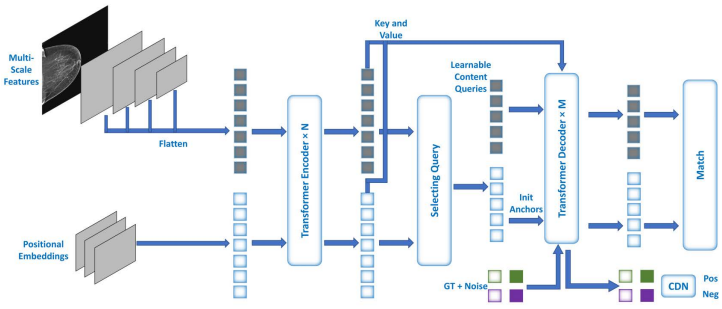

Fig. 3: The architecture of the proposed model for breast cancer detection.

3 Experiments and Results

3.1 Dataset and Setup

In our study, we curated a dataset consisting of 12,476 mammographic images, meticulously processed according to BI-RADS standards, and prepared for machine learning applications by expert radiologists who performed data cleaning and Region of Interest (ROI) extraction. We split this dataset into training 90%(11,228 images) and validation 10% (1,248 images) sets to facilitate a comprehensive training regime. During the fine-tuning stage, image sizes were augmented by 1.5 times their original dimensions to enhance the model's ability to recognize detailed features. Our model's training was optimized using an initial learning rate of 1×10^{-4} and the AdamW[7, 11] optimizer, with a weight decay also set to 1×10^{-4}. For loss functions, we employed L1 and Generalized Intersection over Union (GIOU)[14] losses for box regression, along with focal loss (with $\alpha = 0.25, \gamma = 2$) for classification, aiming to refine the precision of bounding box predictions and improve classification accuracy.

The evaluation of our model's performance was conducted using the Average Precision (AP) metric across various Intersection over Union (IoU) thresholds and object scales, providing insights into the model's predictive accuracy and its capability in handling objects of varying sizes. Implementation was conducted using PyTorch on dual NVIDIA GeForce A40 GPUs with 100GB memory each.

Model	Backbone	Scale	IOU							
			20	30	40	50	60	70	80	Avg
IHRRB-DINO	Resnet50	4	60.0	58.7	56.9	54.8	50.7	37.0	19.0	42.55
	Resnet50	5	58.3	56.9	55.5	53.7	49.5	37.7	21.9	42.04
	Swin-L	4	**63.2**	**62.4**	**61.1**	**59.5**	**56.4**	**46.3**	**22.6**	**46.96**

Table 1: Localization performance at different IoU thresholds

3.2 Localization Performance at Different IoU Thresholds

Table 1 presents a detailed analysis of the IHRRB-DINO model using different backbones across a different range of IoU thresholds, highlighting the ability to localize breast cancers with varying degrees of precision. The model's performance with the Swin-L[10] backbone at a scale of 4 distinguished significantly, achieving the highest average IoU score of 46.96. This is notably superior to the same model utilizing a ResNet50 backbone, where the highest average IoU attained is 42.55. The substantial difference in the scores, particularly at higher IoU thresholds such as 70, 80, and 90, emphasizes the Swin-L backbone's effectiveness in accurately delineating breast masses in mammographic images. For instance, at an IoU threshold of 80, the Swin-L backbone achieves a score of

22.6 compared to only 19.0 and 21.9 with the ResNet50 backbone, illustrating its superior precision in high-risk region identification.

The model's robust performance across a spectrum of IoU thresholds highlights its potential clinical utility. High accuracy in localizing breast masses is crucial in clinical settings, where precise identification of high-risk regions can significantly impact diagnostic decisions and treatment planning. The IHRRB-DINO model, particularly with the Swin-L backbone, appears to offer the needed precision for effective clinical application. as shown in Figure 4

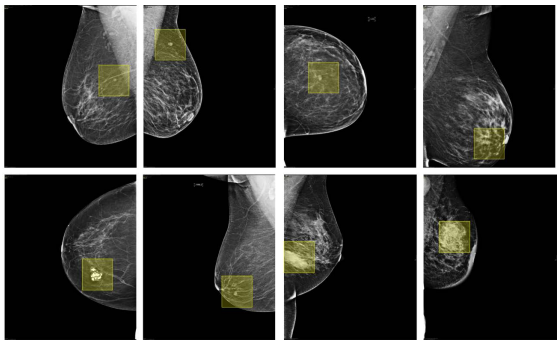

Fig. 4: Examples of Our Proposed IHRRB-DINO Model Output.

3.3 Localization Performance Across Different Models

Our advanced breast cancer detection model, IHRRB-DINO, has undergone rigorous testing using a comprehensively curated dataset. In comparison to prevailing state-of-the-art techniques, including well-established frameworks like ResNet50 and VGG16, our model consistently outperforms them. Even when these traditional methods are enhanced with sophisticated techniques such as Class Activation Mapping (CAM) [23], Heatmap-based Anomaly Segmentation (HAS) [8], Self-Produced Guidance (SPG) [22], Adversarial Discriminative Localization (ADL) [3], and Anomaly Detection with Localization (ACOL)[21].

Table 2 extends the analysis to compare the IHRRB-DINO model with other models like CAM, HAS, ACOL, and ADL, using various backbones. Notably, IHRRB-DINO with a Swin-L backbone significantly outperforms other models, particularly at higher IoU thresholds, which is critical for accurate and reliable cancer detection. For example, at an IoU threshold of 60, the IHRRB-DINO model scores 56.4, surpassing the CAM model with Resnet50 (21.09%), VGG-16 (11.84%), and Inception v3 (5.21%). When we look at how well different models perform at an IoU threshold of 80, it is clear that the IHRRB-DINO model is doing something right. It manages to score 46.96%, while most other models do not even get on the scoreboard. This tells us that the IHRRB-DINO

model is getting better at spotting and pinpointing the location of breast cancers. The IHRRB-DINO model, which uses the Swin-L backbone, is showing some impressive results at higher IoU thresholds. For instance, at an IoU of 60, our model scores 56.4%. This score tells us that our model is good at pinpointing the exact location of breast cancers, which is super important for detecting them accurately. This is a big step up from the next best model, ADL with Resnet50, which only scores 24.64% at the same IoU level. And even when the IoU thresholds get tougher, like 80, IHRRB-DINO still scores 22.6%, showing its precision. In real-world terms, this level of accuracy can make a big difference in catching breast cancer early and planning treatment, which can improve the patient's chances of recovery. The fact that IHRRB-DINO can keep up its high performance at IoU thresholds is really important for doctors. In breast cancer screening, it's just as important to catch small or early-stage tumors as it is to identify larger ones. The models strong performance across different IoU thresholds suggests it could be really useful in a variety of situations, from catching cancer early to dealing with more advanced cases.

Model	Backbone	IOU							
		20	30	40	50	60	70	80	Avg
CAM	Resnet50	58.29	46.68	37.91	29.85	21.09	9.47	1.65	29.28
	VGG-16	54.97	40.75	29.62	21.09	11.84	4.73	1.42	23.49
	Inception v3	48.10	31.75	19.66	9.95	5.21	2.13	0.23	16.72
HAS	Resnet50	18.00	10.66	9.00	6.39	3.79	1.18	0.23	7.03
	VGG-16	25.5	15.43	10.28	5.65	5.01	2.82	0.57	9.32
	Inception v3	2.13	0.94	0.47	0.23	0.23	0.23	0.0	0.61
SPG	Resnet50	28.87	21.72	14.25	9.45	5.40	1.23	0.0	11.56
	VGG-16	48.58	29.89	20.00	9.69	3.15	0.84	0.0	16.02
	Inception v3	30.42	20.13	9.67	5.87	2.14	0.54	0.0	9.82
ACOL	Resnet50	31.56	25.11	19.43	12.55	8.53	4.02	0.94	14.59
	VGG-16	30.46	20.68	13.34	10.04	6.18	2.13	0.56	11.90
	Inception v3	4.97	3.08	1.89	1.18	0.71	0.23	0.23	1.75
ADL	Resnet50	**68.72**	55.45	44.31	33.64	24.64	14.45	5.68	35.27
	VGG-16	43.60	24.17	12.08	5.92	2.13	0.94	0.0	12.70
	Inception v3	33.64	16.82	9.00	3.31	1.18	2.23	2.23	9.77
IHRRB-DINO	Swin-L	63.2	**62.4**	**61.1**	**59.5**	**56.4**	**46.3**	**22.6**	**46.96**

Table 2: Localization performance at different IoU thresholds

4 Conclusion

In conclusion, this study introduces the IHRRB-DINO model, a new approach for the detection and localization of breast masses in mammogram images. Leveraging the advanced capabilities of ViTs, particularly with the Swin-L backbone,

our model surpasses existing methods in high-risk areas in breast tissue. The rigorous testing we conducted shows that IHRRB-DINO outperforms traditional models, especially when it comes to detailed and precise detection at various levels. Key to this model's success is the use of DINO and a novel training strategy, which enhance its ability to identify and analyze complex patterns in mammographic images.

5 Compliance with Ethical Standards

Data used for this study was collected from human subjects who provided their consent. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Qassim Health Cluster in Saudi Arabia (no. 1442-1753964; granted 03 May 2021).

Disclosure of Interests. The authors have no competing interests.

References

1. Arasu, V.A., Habel, L.A., Achacoso, N.S., Buist, D.S., Cord, J.B., Esserman, L.J., Hylton, N.M., Glymour, M.M., Kornak, J., Kushi, L.H., et al.: Comparison of mammography ai algorithms with a clinical risk model for 5-year breast cancer risk prediction: An observational study. Radiology **307**(5), e222733 (2023)
2. Chalfant, J.S., Hoyt, A.C.: Breast density: current knowledge, assessment methods, and clinical implications. Journal of Breast Imaging **4**(4), 357–370 (2022)
3. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2219–2228 (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Gastounioti, A., Conant, E.F., Kontos, D.: Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. Breast cancer research **18**(1), 1–12 (2016)
6. 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)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kumar Singh, K., Jae Lee, Y.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3524–3533 (2017)
9. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–623. Springer (2022)

10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Lotter, W., Diab, A.R., Haslam, B., Kim, J.G., Grisot, G., Wu, E., Wu, K., Onieva, J.O., Boyer, Y., Boxerman, J.L., et al.: Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature Medicine **27**(2), 244–249 (2021)
13. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al.: International evaluation of an ai system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
14. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
15. Shareef, B., Xian, M., Vakanski, A., Wang, H.: Breast ultrasound tumor classification using a hybrid multitask cnn-transformer network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 344–353. Springer (2023)
16. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
17. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., et al.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical image analysis **68**, 101908 (2021)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
19. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE transactions on medical imaging **39**(4), 1184–1194 (2019)
20. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology **292**(1), 60–66 (2019)
21. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1325–1334 (2018)
22. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: Proceedings of the European conference on computer vision (ECCV). pp. 597–613 (2018)
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
24. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)