Exemplar Med-DETR: Toward Generalized and Robust Lesion Detection in Mammogram Images and Beyond

Sheethal Bhat[1,3][0009−0006−2044−5113], Bogdan Georgescu[2], Adarsh Bhandary Panambur[1], Mathias Zinnen[1], Tri-Thien Nguyen[1], Awais Mansoor[2], Karim Khalifa Elbarbary[1], Siming Bayer[1,3], Florin-Cristian Ghesu[3], Sasa Grbic[2], and Andreas Maier[1]

[1] Friedrich-Alexander-Universität, Erlangen, 91058, Germany
[2] Siemens Medical Solutions, Princeton, NJ, 08450, U.S.A
[3] Siemens Healthineers, Erlangen, 91052, Germany
sheethal.bhat@fau.de

Abstract. Detecting abnormalities in medical images poses unique challenges due to differences in feature representations and the intricate relationship between anatomical structures and abnormalities. This is especially evident in mammography, where dense breast tissue can obscure lesions, complicating radiological interpretation. Despite leveraging anatomical and semantic context, existing detection methods struggle to learn effective class-specific features, limiting their applicability across different tasks and imaging modalities. In this work, we introduce Exemplar Med-DETR, a novel multi-modal contrastive detector that enables feature-based detection. It employs cross-attention with inherently derived, intuitive class-specific exemplar features and is trained with an iterative strategy. We achieve state-of-the-art performance across three distinct imaging modalities from four public datasets. On Vietnamese dense breast mammograms, we attain an mAP_{50} of 0.7 for mass detection and 0.55 for calcifications, yielding an absolute improvement of 16% points from previous state-of-the-art. Additionally, a radiologist-supported evaluation of 100 mammograms from an out-of-distribution Chinese cohort demonstrates a twofold gain in lesion detection performance. For chest X-rays and angiography, we achieve an mAP_{50} of 0.25 for mass and 0.37 for stenosis detection, improving results by 4% and 7% points, respectively. These results highlight the potential of our approach to advance robust and generalizable detection systems for medical imaging.

Keywords: Computer-Aided Diagnosis · Lesion Detection · Mammography.

1 Introduction

With the advent of advanced deep learning algorithms, computer-aided diagnosis tools have improved significantly over the past decade [1]. Notably, in the context

of breast cancer screening using mammography, substantial progress has been made in developing artificial intelligence (AI)-supported screening and diagnostic systems [8,9,24]. However, mammography screening sensitivity decreases with higher breast density [27] particularly in populations such as Vietnam and China. Given the increased prevalence of dense breast tissue among women in these populations [4], developing algorithms to address this issue is crucial. Challenges arise not only from the data imbalance across different breast densities [22] leading to algorithmic bias, but also because detection is complicated by dense breast tissue obscuring lesions [27,23].

Recent methods for object detection in the natural image domain such as Faster-RCNN [26], RetinaNet [13], DETR [33], and YOLO [28,25] have produced impressive results in various medical imaging domains. A recent study by Chen et al. [9] implemented a multi-modal image-text EfficientNet-based network along with a RetinaNet detector [13] to achieve state-of-the-art (SOTA) precision for lesion detection in mammograms [22]. On a similar note, Rangarajan et al. [24] examined the detection of lesions in dense breast populations, while Marimuthu et al. [18] investigated the use of imaging and anatomical information for lesion detection using deep learning models. Although current studies achieve impressive results by either using multi-modal or supplementary contextual (spatial and anatomical) information, there remains room for further enhancement. Current methods such as Grounding DINO (GD) [5,15] leverage vision and language information through cross-modal attention and contrastive learning, enabling open-set detection with textual prompts, thus achieving SOTA detection performance on MS-COCO [14] and PASCAL-VOC [10] datasets.

Building on the concept of multi-modal data fusion, we propose **Exemplar Med-DETR**, a novel object detection framework that significantly improves lesion detection in dense breast mammograms. Our approach learns a class-specific representative feature, which is incorporated alongside vision and text inputs in the detection pipeline. These "exemplar" features direct the detection heads to effectively localize lesions in mammograms based on "matching" class-features. The effectiveness of Exemplar Med-DETR (EM-DETR) depends on the contrastive learning scheme between classes. This poses two main challenges: differentiating anatomical features from abnormalities and distinguishing closely related class features. Secondly, anatomies are inherently aligned in medical images compared to natural images, where the object classes can occur at any location. Exploiting these insights, EM-DETR introduces an iterative training scheme that trains the network in stages, contrasting between normal anatomies and regions with lesions. In this study, we perform a comprehensive evaluation of our approach on diverse public datasets and assess the impact of the introduced modules with a range of ablation studies.

Main contributions: 1) We introduce EM-DETR that enables feature-based detection in medical images. This is achieved through an *Exemplar generation* module that extracts class-specific representative embeddings to guide detection. 2) We enhance the contrastive learning pipeline [15] with domain-specific background selection, achieving notable gains in abnormality detection using an

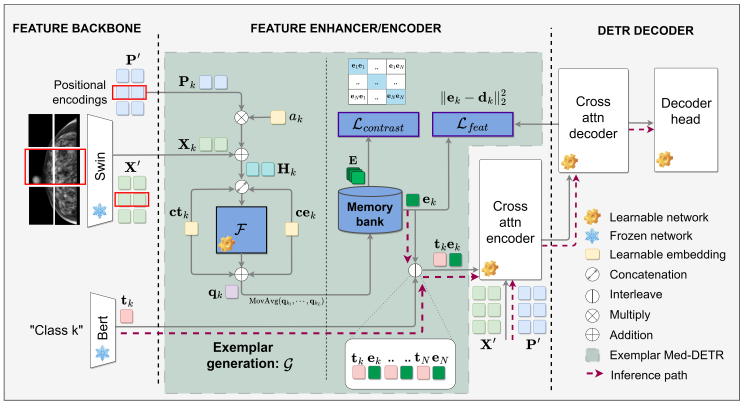

Fig. 1. Overview of EM-DETR. Visual features $\mathbf{X}_k$ & positional embeddings $\mathbf{P}_k$ are extracted from the frozen image backbone based on the k^{th} class location (red). $\mathbf{F}$ uses learnable class-specific embeddings $\mathbf{ct}_k$ & $\mathbf{ce}_k$ to calculate $\mathbf{q}_k$. A moving average of $\mathbf{q}_k$ results in the representative embedding $\mathbf{e}_k$. Interleaving the text embedding $\mathbf{t}_k$ with $\mathbf{e}_k$ enables text-and feature-based detection for each class. $\mathbf{e}_k$ is used in additional $\mathcal{L}_{contrast}$ & $\mathcal{L}_{feat}$ losses to improve the decoder search operations. Inference path in dotted arrows.

iterative training strategy. 3) We demonstrate significant gains in lesion detection in mammography [22] and further evaluate on a public Chinese mammogram dataset [7], comparing with a board-certified radiologist. To also assess generalizability, we extend the method to detect lesions in chest X-rays (CXRs) [21] and stenosis in angiography [20] datasets.

2 Method

Our method is based on multi-modal DETR [3,15] that performs "language guided query selection" through cross-attention between image and text embeddings. We propose to learn class-specific example embeddings or features that can additionally guide detection. These exemplars are computed from the visual features that correspond to the spatial location of the respective classes. Attending to these features enables detection heads to perform a prototype-based search, allowing for an easy expansion to novel classes. Fig. 1 gives an overview of EM-DETR that includes exemplar generation and additional losses. Furthermore, EM-DETR employs an iterative training strategy.

Exemplar Generation $\mathcal{G}$: The Swin transformer [16] image backbone is a multi-scale, shifted window transformer that produces a set of J patch embeddings $\mathbf{X'} = [\mathbf{x_1}, ...\mathbf{x}_J]^T$ for an input image, where each $\mathbf{x}_j \in \mathbb{R}^d$. A similar dimension set of positional encodings is also generated, denoted as $\mathbf{P'} = [\mathbf{p_1}, ...\mathbf{p}_J]^T$. Let $\mathbf{X}_k \subset \mathbf{X'}$ and $\mathbf{P}_k \subset \mathbf{P'}$ represent the set of M tokens that fall within the region of the class k (red).

First, $\mathbf{P}_k$ is scaled by a learnable class-specific parameter $a_k \in \mathbb{R}$ and added to $\mathbf{X}_k$. This produces $\mathbf{H}_k$, mathematically shown as

$$\mathbf{H}_k = \mathbf{X}_k + \mathbf{P}_k \times a_k. \tag{1}$$

a_k modulates the impact of the positional embeddings, thereby influencing the decoder search operation. Additional learnable class-wise token $\mathbf{ct}_k$ and positional $\mathbf{ce}_k$ embeddings are then concatenated with $\mathbf{H}_k$. The simple attention pooling transformer $\mathcal{F}$ processes the concatenated result to learn a single normalized embedding using self-attention mechanism [19]. These class-wise embeddings ($\mathbf{ct}_k$ and $\mathbf{ce}_k$) are also added to the output of $\mathcal{F}$ to produce a class-specific feature embedding $\mathbf{q}_k \in \mathbb{R}^d$ encapsulating the textural and spatial features corresponding to class k. The addition of $\mathbf{ct}_k$ and $\mathbf{ce}_k$ is analogous to the positional embeddings added to the text embeddings in the feature enhancer [15,32]. Mathematically, $\mathbf{q}_k$ is given as

$$\mathbf{q}_k = \mathcal{F}(concat(\mathbf{ct}_k, \mathbf{ce}_k, \mathbf{H}_k)) + \mathbf{ct}_k + \mathbf{ce}_k, \tag{2}$$

where $k \in \{1, \ldots, N\}$ and N is the total number of classes. A moving average is calculated over L samples of $\mathbf{q}_k$, thereby generating the representative embedding $\mathbf{e}_k \in \mathbb{R}^d$. This prevents the decoder from pursuing rapidly changing features during each training iteration. $\mathbf{e}_k$ is stored in a *Memory bank* as a prototype feature embedding that helps prevent catastrophic forgetting. The memory bank of saved feature representations, or exemplars $\mathbf{e}_k$, are also used for inference.

In parallel, the frozen text encoder processes text prompts and produces embeddings $\mathbf{t}_k$ which are interleaved with the corresponding $\mathbf{e}_k$ and passed downstream to the encoder-decoder pipeline [3,15]. Text prompts are literal class names (e.g. "mass", "stenosis", "background"). The decoder predicts the k^{th} class location by processing the cross-attention encodings of the text and representative visual features, $\mathbf{t}_k$ and $\mathbf{e}_k$ with the input image embeddings, $\mathbf{X}'$ and $\mathbf{P}'$.

Additional Losses $\mathcal{L}$: In addition to the original DETR loss terms $\mathcal{L}_{\text{bbox}}$, $\mathcal{L}_{\text{IoU}}$, and $\mathcal{L}_{\text{classify}}$ in [15] we introduce two additional loss functions to improve the robustness of EM-DETR. A cosine similarity *contrastive feature loss* $\mathcal{L}_{contrast}$, is applied on $E = [\mathbf{e}_1 ... \mathbf{e}_N]^T$ of all $1 \leq k \leq N$ classes. $\mathcal{L}_{contrast}$ ensures that all class representative embeddings remain orthogonal to each other in the latent space, promoting distinct and separable representations [30]. A L_2 *feature loss*, $\mathcal{L}_{feat}$, is applied between $\mathbf{e}_k$ and the decoder's top proposal $\mathbf{d}_k$ for class k ensuring that the model predicts a consistent latent representation for each class [2,32]. This approach helps maintain class-specific embeddings over time while stabilizing the training process and is empirically observed to improve our detection results.

Iterative training strategy: While the trained decoder effectively learns the features of these classes, it faces challenges in distinguishing normal anatomical structures. Therefore, we propose a multi-stage iterative learning approach. Stage I involves training the proposed model with all annotations, while Stage II refines the weights through a per-class background-versus-foreground detection

Table 1. Results for lesion detection on VinDR-Mammo [22] & CMMD [7] test datasets. The std dev. is $< \pm 0.005$ for all runs. $*$ indicates a statistical significance of $p < 0.0001$ compared to the baseline. † for Swin architecture.

Method	Mass[22]		Calcification[22]		CMMD Mass[7]
	$\mathbf{mAP_{50}}$	Recall	$\mathbf{mAP_{50}}$	Recall	TP rate
YOLOV3† [28]	0.52	0.35	0.11	0.15	-
Faster-RCNN† [26]	0.54	0.44	0.22	0.26	-
RetinaNet† [13]	0.58	0.50	0.43	0.36	-
MammoCLIP [9]	0.58	-	0.35	-	-
GD† [15] (Baseline)	0.48	0.61	0.45	0.44	0.32
EM-DETR† (Ours)	**0.70**$*$	**0.66**$*$	**0.55**$*$	**0.50**$*$	**0.67**$*$

task. The background annotations are generated based on the dataset: for mammogram and CXR, random boxes are sampled from normal images. Moreover, in CXR, a pool of training lesion locations is used so that anatomical priors are learned. In contrast, stenosis backgrounds are selected from outside the annotated regions. In an additional Stage III, we denote previously detected False Positive (FP) regions as background classes to further refine the network.

3 Experimental setup

Data description: We utilize the VinDR-Mammo dataset [22] for lesion detection tasks involving mass and calcification. It comprises 16,000 training and 4,000 full-resolution test images from patients in Vietnam, with bounding box (bbox) annotations. The dataset contains a high proportion of dense breast tissue (approx. 90% [22]), making mass detection challenging for both radiologists and AI models. For a fair comparison, the data set is created according to the current SOTA MammoCLIP [9]. Furthermore, we utilize a random subset of 100 images from the Chinese mammogram dataset CMMD [7] to evaluate the lesion detection task. Due to the absence of bbox annotations in this dataset, a board-certified radiologist identified the centers of the lesion region for groundtruth. All CMMD images contain lesions, which impacts labeling. Full-resolution data preprocessing involves cropping the background [23]. We extend the validation study to additional domains to assess our method's robustness and generalizability. VinDR-CXR [21] is a dataset of 15,000 training and 3,000 test postero-anterior (PA) full-resolution images with bbox annotations. We evaluate for a comparable task of nodules and mass detection with a test set created as in [31]. To assess the model with a dissimilar objective, we investigate stenosis detection utilizing the ARCADE [20] angiography dataset. This dataset [20] contains 1000 training and 300 test images. The dataset contains labeling noise where stenosis regions outside the main vessel tree are not annotated in some images, thus minimizing observed gains. These are public datasets and we ensure no patient overlap between training and test sets in all experiments.

Table 2. Results for lesion detection on VinDR-CXR [21] & stenosis detection on ARCADE test datasets [20]. The std dev. is $< \pm 0.01$ for all runs. $*$ & $**$ indicates a statistical significance of $p < 0.0001$ & $p < 0.001$, respectively compared to the baseline.

Category	Method	$\mathbf{mAP_{50}}$	Recall
Nodules/Mass CXR [21]	SAR CNN [12]	0.07	-
	EARL [11]	0.14	-
	DualAttnNet [29]	0.14	-
	Multi-scale Location Aware Detector [31]	0.18	-
	YOLOV7-X [17]	0.21	-
	GD (Baseline) [15]	0.14	0.30
	EM-DETR (Ours)	**0.25**$*$	**0.33**
Stenosis [20]	DINO-DETR [33]	0.23	**0.53**
	YOLO [25]	0.25	0.18
	GD (Baseline) [15]	0.30	0.37
	EM-DETR (Ours)	**0.37**$**$	0.42

Experimental details: The experiments were run on a single node with four 40 GB A100 GPUs, using a learning rate of 0.0008 with a linear scheduler under the MMDetection [6] framework. An average of 5 runs is recorded. The image and text backbones were frozen. $\mathcal{F}$ is designed as a simple 2-head, 4-layer transformer. The moving average is computed over $L = 200$ exemplars. We initially train our model with all available annotations (Stage I) and subsequently refine it by focusing on each class individually (Stage II). We utilize 8 randomly selected background regions from normal images to ensure coverage of the entire image. At the final stage, we further finetune the model with the top 8 misclassified regions (FP) set as background (Stage III). The mean average precision at 50% IoU (mAP_{50}) is reported with an average of 50 to 95% IoU recall values.

4 Results and Discussion

Tables 1 and 2 show the results for lesion detection on mammography, and for lesion and stenosis detection on CXRs and angiography images, respectively. Fig. 2 and 3 illustrate exemplary detection results from each dataset.

Our method achieves SOTA results for both mass and calcification detection on VinDR-Mammo dataset [22] as seen in Table. 1. A significant improvement of 12% mAP_{50} is observed in mass detection, compared to the previous SOTA. As seen in Fig. 2(a), the prediction closely aligns with the groundtruth, even within extremely dense breast tissue. Similarly, an increase of 20% w.r.t. SOTA is observed in the calcification detection results. Despite significant gains, we investigate cases with lower mAP_{50}. We find that calcifications commonly co-occur with mass and are annotated with a common bbox. Our method relies on feature matching and these annotations influence our results, leading to predictions that do not precisely match the groundtruth as seen in Fig. 2(b). In the first calcification image, the model accurately predicts calcification regions

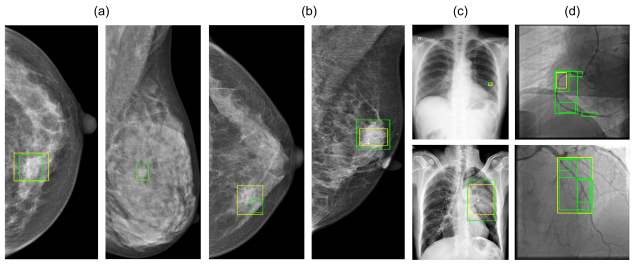

Fig. 2. Detection results of (a) Mass [22], (b) Calcification [22], (c) Nodule [21], & (d) Stenosis [20]. Predictions for the second mass image coincide with groundtruth. Stenosis images show the top 5 predictions, while mammogram & CXR images show the top 1. groundtruth (yellow) & predictions (green). Fig(s) resized for presentation.

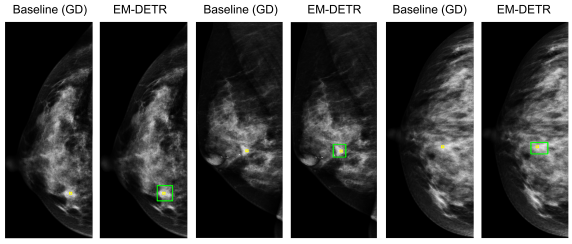

Fig. 3. Example images that compare groundtruth in yellow with predictions (green) between GD & EM-DETR on CMMD [7] images. Through "feature-based" search EM-DETR reliably locates the obscure mass regions.

without including the larger mass. However, in the second image, it includes the mass within the predicted region due to inconsistent feature extraction in $\mathcal{G}$.

To evaluate the out-of-distribution (OOD) performance of mass detection, we use the model previously trained on the mass images of Vindr-Mammo [22] and test it on CMMD [7]. Due to the absence of precise bbox annotations and based on practical clinical relevance, we consider a True Positive (TP) rate as our evaluation metric. A detection is considered a TP, if the groundtruth center is within the highest score predicted box above a threshold of 0.1. With this criteria, we achieve an absolute increase of 35% compared to the baseline as denoted in Table. 1. In Fig. 3, we observe three instances of mass detection between the baseline and EM-DETR along with the groundtruth marked in yellow. For this test, we use the exemplars stored in the memory bank, computed during training for mass detection on VinDR-Mammo dataset. We observe EM-DETR effectively identifies the abnormal regions in OOD dense mammograms. The striking improvements stem from the model's capacity to infer the salient dataset agnostic features of masses. Additional CMMD results are in supplementary file.

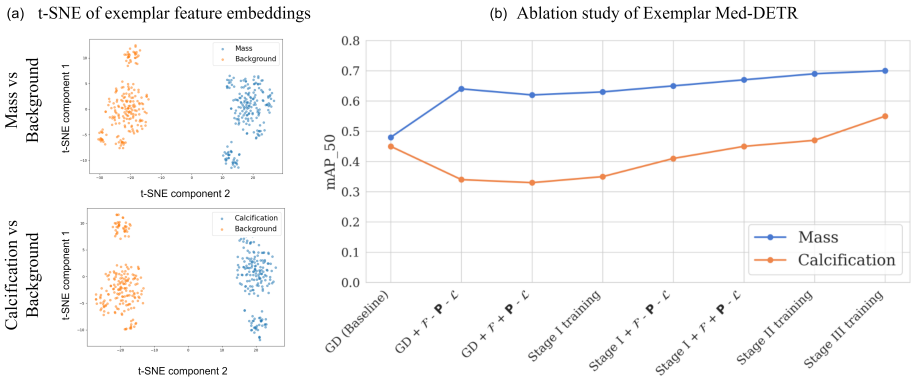

Fig. 4. (a) t-SNE plots of mass & calcification vs. their background memory bank embeddings. (b) Mass & calcification mAP_{50} for different model configurations and training stages. GD is baseline, $\mathcal{F}$, $\mathbf{P}$ as in Sec. 2. $\mathcal{L}$ is $\mathcal{L}_{contrast} + \mathcal{L}_{feat}$ (Sec. 2). Stages as in Sec. 2. Stage I, II & III includes $\mathcal{F}, \mathbf{P}$ & $\mathcal{L}$.

As our method effectively contrasts between normal and pathological anatomy, we also assess nodule and mass detections on a CXR dataset [21]. We achieve a gain of 4% resulting in a new SOTAof 0.25 mAP_{50} as seen in Table. 2. Fig. 2(c) presents examples of nodule detections.

The generalizability of our method is further demonstrated in stenosis detection [20] which differs from the previous tasks, resulting in a 7% improvement in mAP_{50} attaining SOTA. The results of stenosis detection are also depicted in Table. 2. The decoder learning methodology relies on contrasting the proposed hypothesis boxes, and the presence of labeling noise—stemming from stenosis not being annotated in secondary vessels—leads to improvements that fall short of expectations. Moreover, multiple smaller hypothesis boxes that follow the impacted vessel structure are predicted within the larger groundtruth region, also leading to lower precision scores. Fig. 2(d) displays multiple predicted stenosis regions that are within the groundtruth annotations as well as an example of an absent groundtruth annotation.

Ablation studies: Fig. 4(a) presents the t-SNE plots of exemplar features in the case of mass and calcification vs. their backgrounds. The exemplars are observed to be well separated, thus ensuring the decoder searches for discriminative class features. Fig. 4(b) shows the impact of various modules of EM-DETR for mass and calcification across different model configurations and training stages. Starting from the baseline GD, the exemplar generation module is added without and with positional encodings ($\mathbf{P}$ in Sec. 2), and the introduced loss terms ($\mathcal{L}$ in Sec. 2). To provide additional insight, the performance gain is presented through the different stages of training (Sec. 2). The progressive integration of various modules in EM-DETR leads to consistent improvements in mass detection, as evidenced by the increasing mAP_{50}. On the other hand, we observe that

the calcification results initially decrease with the introduction of $\mathcal{F}$, to improve later after Stage II training. This is attributed to the groundtruth incorporating identical annotations for mass and calcification in images featuring both findings.

5 Conclusion

In this study, we demonstrate that EM-DETR efficiently performs detection in various challenging tasks. It achieves SOTA performance through "feature matching" and domain adaptive training. The method ensures the decoder is primed to search based on "exemplar" features, enabling a powerful stage I model that may expand easily to novel classes. Future work will investigate the use of EM-DETR for medical detection foundation models and few-shot detection to reduce annotation cost.

Acknowledgments. This study was funded by Siemens-Healthineers, U.S.A

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bhat, S., Mansoor, A., Georgescu, B., Panambur, A.B., Ghesu, F.C., Islam, S., Packhäuser, K., Rodríguez-Salas, D., Grbic, S., Maier, A.: AUCReshaping: Improved sensitivity at high-specificity. Scientific Reports **13**(1), 21097 (2023) 1
2. Bulat, A., Guerrero, R., Martinez, B., Tzimiropoulos, G.: FS-DETR: Few-shot detection transformer with prompting and without re-training. In: ICCV. pp. 11793–11802 (2023) 4
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020) 3, 4
4. del Carmen, M.G., Halpern, E.F., Kopans, D.B., Moy, B., Moore, R.H., Goss, P.E., Hughes, K.S.: Mammographic breast density and race. American Journal of Roentgenology **188**(4), 1147–1150 (2007) 2
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS **33**, 9912–9924 (2020) 2
6. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 6
7. Cui, C., Li, L., Cai, H., Fan, Z., Zhang, L., Dan, T., Li, J., Wang, J.: The Chinese Mammography Database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast. The Cancer Imaging Archive (2021). https://doi.org/10.7937/tcia.eqde-4b16 3, 5, 7
8. Díaz, O., Rodríguez, A., Sechopoulos, I.: AI for breast cancer detection: Technology, challenges, and prospects. Eur. J. Radiol. p. 111457 (2024) 2
9. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) MICCAI. pp. 632–642. Springer Nature Switzerland, Cham (2024) 2, 5

10. Hoiem, D., Divvala, S.K., Hays, J.H.: Pascal VOC 2008 challenge. World Literature Today **24**(1), 1–4 (2009) 2
11. Le, K.H., Tran, T.V., Pham, H.H., Nguyen, H.T., Le, T.T., Nguyen, H.Q.: Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. IEEE Access **11**, 14105–14114 (2023) 6
12. Lin, C., Huang, Y., Wang, W., Feng, S., Huang, M.: Lesion detection of chest X-Ray based on scalable attention residual CNN. Math Biosci Eng **20**(20), 1730–49 (2023) 6
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 2, 5
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 2
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024) 2, 3, 4, 5, 6
16. 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: ICCV. pp. 10012–10022 (2021) 3
17. Luo, J., Wang, S., Wang, Q., Liu, S.: A Lung Lesion Detection Algorithm Based on YOLOv7 and Self-Attention Mechanism. In: CCC. pp. 8786–8791 (2023) 6
18. Marimuthu, T., Rajan, V.A., Londhe, G.V., Logeshwaran, J.: Deep Learning for Automated Lesion Detection in Mammography. In: ICIDeA. pp. 383–388. IEEE (2023) 2
19. Marin, D., Chang, J.H.R., Ranjan, A., Prabhu, A., Rastegari, M., Tuzel, O.: Token pooling in vision transformers for image classification. In: Proceedings of the IEEE/CVF WACV. pp. 12–21 (2023) 4
20. Maxim Popov, A., et al.: ARCADE: Automatic Region-based Coronary Artery Disease diagnostics using x-ray angiography imagEs Dataset Phase (2023) 3, 5, 6, 7, 8
21. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022) 3, 5, 6, 7, 8
22. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V.: VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. medRxiv (2022) 2, 3, 5, 6, 7
23. Panambur, A.B., Yu, H., Bhat, S., Madhu, P., Bayer, S., Maier, A.: Attention-guided Erasing: Novel Augmentation Method for Enhancing Downstream Breast Density Classification. In: BVM Workshop. pp. 13–18. Springer (2024) 2, 5
24. Rangarajan, K., Aggarwal, P., Gupta, D.K., Dhanakshirur, R., Baby, A., Pal, C., Gupta, A.K., Hari, S., Banerjee, S., Arora, C.: Deep learning for detection of iso-dense, obscure masses in mammographically dense breasts. European Radiology **33**(11), 8112–8121 (2023) 2
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016) 2, 6
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on PAMI **39**(6), 1137–1149 (2016) 2, 5

27. Sickles, E.A., D'Orsi, C.J., Bassett, L.W., et al.: ACR BI-RADS Mammography. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System, pp. 121–140. Reston, VA, American College of Radiology (2013) 2
28. Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z.: Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Computers and electronics in agriculture **157**, 417–426 (2019) 2, 5
29. Xu, Q., Duan, W.: DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays. Computers in Biology and Medicine **168**, 107742 (2024) 6
30. Yang, F., Wu, K., Zhang, S., Jiang, G., Liu, Y., Zheng, F., Zhang, W., Wang, C., Zeng, L.: Class-aware contrastive semi-supervised learning. In: CVPR. pp. 14421–14430 (2022) 4
31. Yuan, Y., Liu, L., Yang, X., Liu, L., Huang, Q.: Multi-scale Lesion Feature Fusion and Location-Aware for Chest Multi-disease Detection. JIIM pp. 1–16 (2024) 5, 6
32. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. IEEE transactions on PAMI **45**(11), 12832–12843 (2022) 4
33. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022) 2, 6