Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-023-04517-9

ORIGINAL RESEARCH

Transformer-based mass detection in digital mammograms

Amparo S. Betancourt Tarifa[1] · Claudio Marrocco[1] · Mario Molinara[1] · Francesco Tortorella[2] · Alessandro Bria[1]

Received: 13 September 2022 / Accepted: 2 January 2023

Abstract
In the last decade, Convolutional Neural Networks (CNNs) have been the de facto approach for automated medical image detection. Recently, Vision Transformers have emerged in computer vision as an alternative to CNNs. Specifically, the Shifted Window (Swin) Transformer is a general-purpose backbone that learns attention-based hierarchical features and achieves state-of-the-art performances in a variety of vision tasks. In this work, for the first time, we design and experiment transformer-based models for mass detection in digital mammograms leveraging Swin transformer as a backbone multiscale feature extractor. Experiments on the largest publicly available mammography image database OMI-DB yield a True Positive Rate (TPR) of 75.7% at 0.1 False Positives per Image (FPpI) for the best transformer model, with 2.5% TPR improvement over its convolutional counterpart and a massive 7.4% TPR over the state-of-the-art. We also combine transformer- and convolution-based detectors with weighted box fusion, achieving an additional 2.4% TPR improvement reaching 78.1% TPR at 0.1 FPpI.

Keywords Mammography · Breast cancer · Deep learning · Transformers · Computer aided detection

1 Introduction

In 2020, female breast cancer was the most commonly diagnosed cancer, with an estimated 2.3 million new cases (11.7%) and 685,000 deaths globally, representing the leading cause of cancer deaths among women (Sung et al. 2021). Mammography, despite known limitations, is still the most commonly used imaging technique for early detection of breast cancer in women over the age of 40 (CDC 2022). Standard mammographic screening consists of mediolateral oblique (MLO) and craniocaudal (CC) low-energy x-ray 2D projection images of each breast to detect suspicious lesions like masses, which appear with characteristic shape and contour (see Fig. 1). Studies have shown a mortality reduction of about 40% after the implementation of mammography screening (Sankatsing et al. 2017). However, since mass detection is a manual and difficult process, a significant proportion of breast masses are missed (Wang et al. 2014). Computer-Aided Detection (CADe) systems based on Artificial Intelligence (AI) technologies can assist radiologists in the detection and localization of masses or other anomalies. Studies on the efficiency of using CADe systems as second opinion reveal that they can benefit even experienced radiologists by increasing their sensitivity from 77 to 85% and beginner radiologists from 62 to 86% (Balleyguier et al. 2005). However, the use of these systems and in general of AI in medicine can introduce social and ethical challenges to security, privacy, and human rights, which deserve attention and investigation (Rajpurkar et al. 2022; Johnson et al. 2021).

With significant advancements in the development of deep learning technologies over the last ten years, CADe systems have been predominantly built using

✉ Alessandro Bria
a.bria@unicas.it

Amparo S. Betancourt Tarifa
ampi.betancourt@gmail.com

Claudio Marrocco
c.marrocco@unicas.it

Mario Molinara
m.molinara@unicas.it

Francesco Tortorella
ftortorella@unisa.it

[1] Department of Electrical and Information Engineering, University of Cassino and Southern Latium, Via G. Di Biasio 43, Cassino 03043, Italy

[2] Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II 132, Fisciano 84084, Italy

Published online: 25 January 2023

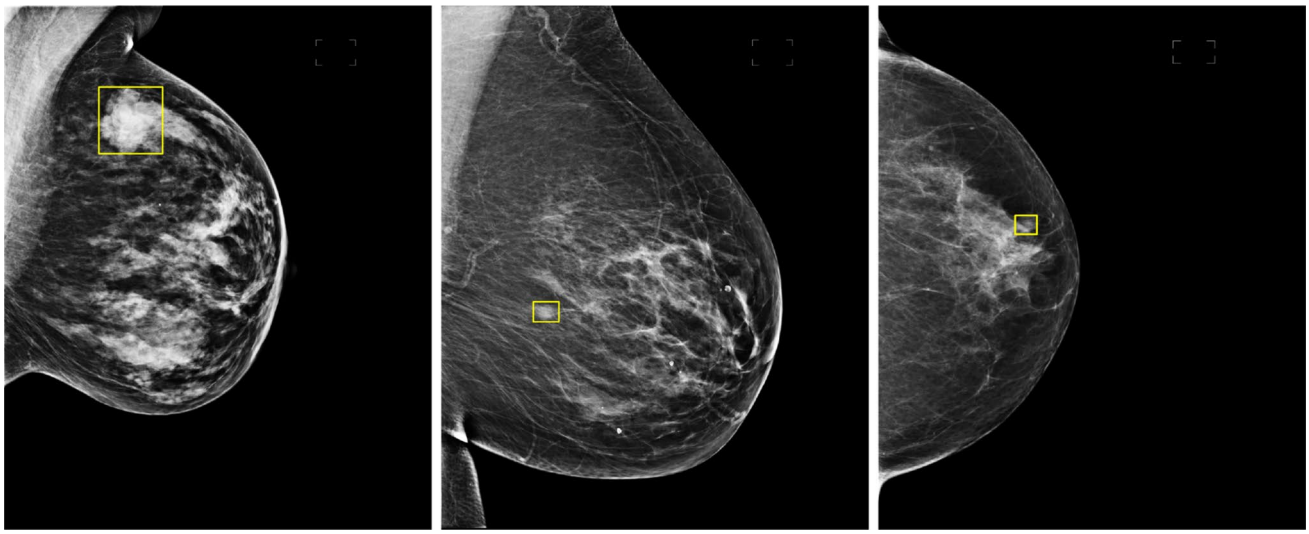

Fig. 1 Mammogram images from OMI-DB containing masses marked with bounding boxes. Conspicuity levels decreasing from left to right: obvious, subtle, occult

Convolutional Neural Networks (CNNs) (Malliori and Pallikarakis 2022). However, the tremendous success of transformer architectures in Natural Language Processing (NLP) has led researchers to explore its adaptation to computer vision where it has emerged as a viable alternative to CNNs after the inception of Vision Transformers (ViT) in the seminal work of Dosovitskiy et al. (2020). Valanarasu et al. (2021) concluded that ViT's self-attention processes are more effective than conventional CNNs at capturing both local and distant visual dependencies. Therefore, the medical imaging community has witnessed an exponential growth of the number of transformer-based approaches focusing on classification and segmentation tasks (Shamshad et al. 2022), whereas detection methods still rely on convolutional backbones for feature extraction [e.g. DETR (Zhu et al. 2020)].

The goal of this study is to explore the use of transformers as backbone feature extractors for mass detection in mammography, and to compare and combine them with their convolutional counterparts. The key contributions of this paper are summarized as follows:

1. proposal of a mass detection framework leveraging a hierarchical transformer as a backbone multiscale feature extractor;
2. comparison of transformer models with their convolutional counterparts;
3. combination of detection predictions from transformer models, their convolutional counterparts, and both;
4. comparison with state-of-the-art on the largest publicly available mammography image database OMI-DB.

2 Related Work

From the early 1990s, academic and business circles have set off a research to develop computer-aided detection and diagnosis technologies that can act as a second opinion or helper for radiologists. This research began with pure image processing techniques (Heath et al. 2000; Petrick et al. 1996; Te Brake and Karssemeijer 1999), including approaches to select regions of interest like breast-air (Méndez et al. 1996; Petrick et al. 1999) and pectoral muscle (Ferrari et al. 2004; Kwok et al. 2004; Molinara et al. 2013) segmentation, and later moved to Machine Learning (ML) based on handcrafted features. Ke et al. (2010) used bilateral comparison to detect the mass and the center of region of interest (ROI), followed by the calculation of fractal dimension and two-dimensional entropy as the texture features. Lastly, the type of ROI was classified by Support Vector Machine (SVM) as mass or normal region. The method achieved a sensitivity of 85.11% at 1.44 false positives per image (FPpI), in a total of 106 mammograms. Patel et al. (2019) presented an effective approach to detect masses in breast images using Modified Histogram based Adaptive Thresholding (MHAT) method, testing it on more than 100 mammograms obtaining a TPR of 98.3% at 0.78 FPpI. Years later, in the work of Mughal et al. (2017), texture features were also used along with color features to detect and classify masses. Methods such as region growing were also proposed as in Punitha et al. (2018). This work used an optimized region growing technique where the initial seed points and thresholds were optimally generated using a swarm optimization technique called Dragon

Fly Optimization. Features were then extracted from the detected masses and inputted to a feed-forward neural network for classification. The approach achieved a sensitivity of 98.1% with specificity of 97.8%, using 300 images from the Digital Database for Screening Mammography (DDSM).

Lbachir et al. (2021) proposed a full CAD system for mass detection and diagnosis applying ML techniques. Firstly the image went through a preprocessing step for image enhancement and noise removal, followed by the segmentation of abnormalities using their proposed Histogram regions analysis-based K-means. False positives were then reduced using texture and shape features inputted to a bagged trees classifier and the abnormalities finally classified by a SVM as malignant or benign. The system was able to achieve a 90.85% TPR at 0.65 FPpI and a 90.44% classification accuracy on the CBIS-DDSM dataset.

Recently, deep learning models employed in computer vision such as convolutional backbones [ResNet (He et al. 2016a), DenseNet (Huang et al. 2017), etc.] and convolutional anchor-based object detection heads [Faster R-CNN (Ren et al. 2015), RetinaNet (Lin et al. 2017), YOLO (Redmon et al. 2016a), etc.], have contributed to significant improvements in the performance of CADe systems. Ribli et al. (2018) used Fast R-CNN on a subset of the INbreast database to detect and classify malignant and benign lesions, achieving 90.0% TPR at 0.30 FPpI. Agarwal et al. (2020) presented for the first time the benchmark of the performance of deep learning on the largest publicly available mammography image database OMI-DB (Halling-Brown et al. 2020). In their work, a framework based on Faster R-CNN achieved 87.0% TPR at 0.84 FPpI on a subset of 7245 images acquired with Hologic scanners. Cao et al. (2021) proposed an anchor-free convolutional model for mass detection along with a new data augmentation technique to overcome overfitting based on local elastic deformation which enhanced the performance of their model at the cost of slower computational speed. This approach leveraged an enhanced, anchor-free version of RetinaNet named Feature Selective Anchor-Free (FSAF) (Zhu et al. 2019) previously proposed in computer vision research, achieving 93.0% TPR at 0.50 FPpI on INbreast. Yu et al. (2022) proposed a hybrid framework which relied on traditional image processing and deep learning techniques. The framework consisted of three main modules: (i) pre-processing, where an improved Deeplabv3+ model for pectoral muscle removal was employed; (ii) a multiple-level thresholding segmentation method to extract candidate mass patches; and (iii) classification into breast mass and breast tissue background by trained deep learning models. On CBIS-DDSM, the method achieved a TPR of 87% at 2.86 FPpI, whereas it reached 96% on INbreast with an FPpI of 1.29. In the work of Yan et al. (2021), a multitasking framework for breast mass detection that combined CC and MLO mammograms was proposed. An image detection pipeline based on YOLOv3 region proposals was followed by a Siamese network that integrated patch level mass vs. non mass classification and dual view mass matching. This approach was evaluated on the INbreast dataset reaching a 96% TPR at 0.26 FPpI. Aly et al. (2021) proposed an end-to-end CADe system based on YOLOv3 with k-means generated anchors, which is an improved version of the network proposed by Redmon et al. (2016b) and achieved 92% TPR at 0.086 FPpI on the INbreast dataset. Aiming to mitigate the presence of excessive negative boxes in anchor-based detection techniques, an anchor-free YOLOv3 was presented by Zhang et al. (2022). This method achieved a 95% TPR at 1.7 FPpI on the INbreast dataset outperforming the traditional YOLOv3 network. Su et al. (2022) proposed a double shot model that combined YOLOV5 and Local–Global (LOGO) transformers for mass detection and segmentation, using the first to place and crop the breast mass in mammograms followed by the segmentation performed with a gated axial-attention mechanism and LOGO training strategy. The proposed model was evaluated on two independent mammography datasets (CBIS-DDSM and INbreast) where it achieved a TPR of 95.7% and mean average precision of 65.0%.

Approaches using transformers have just begun to appear in the literature of mass detection and breast imaging in general. Kamran et al. (2022) introduced Swin Spatial Feature Transformer Network (Swin-SFTNet), a U-net-shaped transformer-based architecture, that outperforms state-of-the-art architectures in breast mammography-based micro-mass segmentation. This method was evaluated on three publicly available datasets, achieving a segmentation dice improvement over the state of the art by 3.10%, 3.81%, and 3.13% on CBIS-DDSM, INbreast, and CBIS, respectively. Chen et al. (2022) proposed a Multi-view Vision Transformer architecture able to separately learn patch relationships between four mammograms acquired from two-views (CC/MLO) of two-side (right/left) breasts, by employing local and global transformer blocks. Their proposed transformer-based model was evaluated on a private dataset including 470 malignant and 479 benign cases with an area under the ROC curve of 0.818, statistically significantly outperforming the state-of-the-art multi-view CNNs.

Table 1 Counts of the dataset used in this work

	Patients	Images	Abnormal	Normal
OMI-H-MD	1945	7626	3526	4100
Training	1361	5339	2478	2861
Validation	195	766	349	417
Test	389	1521	699	822

3 Materials

3.1 Dataset

OMI-DB is an extensive mammography image database composed of more than 2.5 million images from over 170,000 women that were collected from three UK breast screening centres (Halling-Brown et al. 2020). It provides digital mammograms in DICOM format from detected cancers along with normal and benign screening cases. The database contains images from different scanner manufacturers such as Hologic Inc., Siemens, Philips, and General Electric Medical Systems. For this study, images from Hologic Inc. scanners were selected as they represented the vast majority of images in the dataset. From here, we extracted the largest possible subset suited for mass detection, hereafter referred to as OMI-H-MD. It consists of 7, 626 DICOM 'for presentation' screening mammograms from 1, 945 patients, with both detected masses (*positive* or *abnormal* images) and without any abnormality (*negative* or *normal* images), see Table 1. Careful visual inspection of all the selected images was performed, ensuring to discard images with artifacts or unwanted objects such as implants, marker clips or bands across the image. The following criteria were adopted for selecting normal images: (i) when multiple studies for same patient were available, only the images from the first study were considered; (ii) normal images belonging to a patient with abnormalities, even if they were present on a different breast or in a different study, were not considered; (iii) normal images belonging to a study with only one breast or one view were not considered routine screening and thus they were discarded.

3.2 Data preparation

The OMI-H-MD dataset was divided into training, validation and test sets on patient basis to guarantee that images from a particular case belonged exclusively to one of the three subsets. The division was performed as in Agarwal et al. (2020) on a 70-10-20 ratio (see Table 1). It is worth mentioning that even though the amount of images and patients in our dataset is very close to that of their work, it is not an exact match.

All mammograms were converted to 8-bit three-channel PNG images using channel-replication which is widely adopted in medical imaging for finetuning networks pretrained on natural images (Tajbakhsh et al. 2016; Zhou et al. 2017). Pixel resolution was downsampled at 200μm for faster processing like in Agarwal et al. (2020). In addition, useful breast areas were cropped by applying triangle binarization followed by largest connected component selection. All crops were visually checked and manually corrected when necessary ($< 1\%$ of cases).

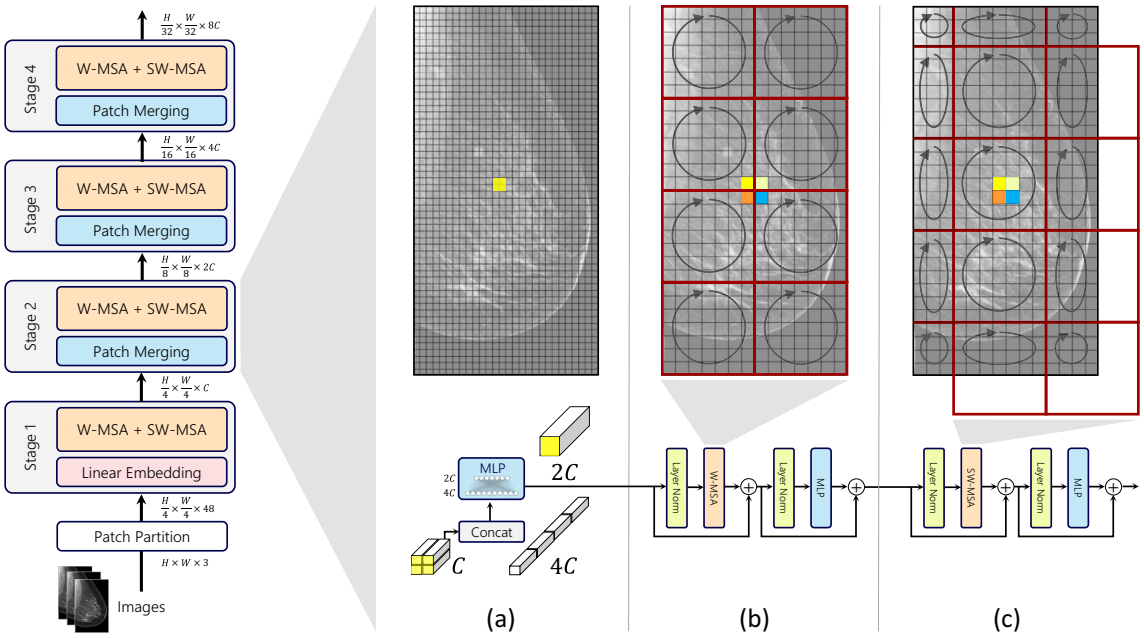

Fig. 2 Swin Transformer backbone architecture for multiscale feature extraction on mammograms with detailed stage substeps: **a** patch merging; **b** Swin transformer block with window self attention; **c** Swin transformer block with shifted window self attention allowing the four colored patches previously belonging to four distinct windows to attend each other

Table 2 Swin architecture variants

Model	C	layers $\{n_i\}$	#heads	#param
Swin-T	96	$\{2, 2, 6, 2\}$	3×4	29M
Swin-S	96	$\{2, 2, 18, 2\}$	3×4	50M
Swin-B	128	$\{2, 2, 18, 2\}$	4×4	88M
Swin-L	192	$\{2, 2, 18, 2\}$	6×4	197M

4 Methodology

We propose using a general-purpose hierarchical vision transformer backbone as multiscale feature extractor, the Shifted Window (Swin) Transformer, initialized with ImageNet pre-trained weights. Then, we combine it with two object detection heads initialized with COCO pre-trained weights and we fine-tune the two architectures on our task. Finally, we propose fusing the predictions of the two detectors, their convolutional-backbone counterparts, and both to investigate whether convolutional-based and transformer-based detectors can complement each other and provide an overall boosted detection performance. In the sections below are described: (i) the ImageNet and COCO datasets on which the backbone and detection heads were pretrained; (ii) the selected transformer backbone; (iii) the object detection heads; and (iv) the bounding boxes fusion approach.

4.1 ImageNet and COCO datasets

ImageNet (Deng et al. 2009) is a large visual database consisting of more than 14 M images manually annotated according to nearly 20,000 categories. Since the seminal work of Krizhevsky et al. (2017) who introduced the ImageNet 2012 Challenge winner AlexNet, all major proposed deep learning backbones are trained on ImageNet and are publicly and freely available for fine-tuning.

Microsoft Common Objects in Context (COCO) is a large-scale image dataset that gathers 328,000 images of complex everyday scenes containing common objects in their natural context. The dataset contains 91 object categories along with bounding box and segmentation mask annotations which are useful to train deep learning models for object detection and instance segmentation.

4.2 Swin transformer

Swin (**S**hifted **win**dow) transformer is a hierarchical vision transformer capable of serving as a general purpose backbone for computer vision (Liu et al. 2021b). Previous vision transformers like ViT (Dosovitskiy et al. 2020) were based on global self-attention between non overlapping, medium-sized (e.g. 16×16 pixels) image patches at a fixed scale, which is unsuitable for high resolution images and dense tasks like image segmentation and detection, being also computationally limited by its quadratic complexity. In contrast, Swin uses a window-based approach combined with window-shifting at various scales that limits the computation of self attention among small (4×4 pixels) patches within nonoverlapping windows while also allowing cross-window connection. This yields linear complexity to image size and makes Swin suitable for dense vision tasks as well. To this date, Swin and its variants are the backbone architectures of state-of-the-art image classification (Liu et al. 2021a), semantic segmentation (Liu et al. 2021a; Wei et al. 2022), instance segmentation (Li et al. 2022), and object detection (Liu et al. 2021a; Zhang et al. 2022) methods. All the above-mentioned aspects, and in particular the capability to extract hierarchical multiscale attention features achieving state-of-the-art performances in dense computer vision tasks, make Swin the optimal backbone for our transformer-based mass detection framework.

4.2.1 Overall architecture

The mammogram is firstly split into non overlapping patches of 4×4 pixels. The raw pixel values of each patch are concatenated into feature vectors of dimension $4 \times 4 \times 3 = 48$ and projected to an embedding of size C by a linear layer. These embedded patches are processed by subsequent stages $i = 1, 2, ..., N$ that alternate local self-attentions within windows and patch merging to achieve multiscale hierarchical feature extraction. In the following, these two fundamental blocks are detailed. An overview of the entire architecture is presented in Fig. 2.

Shifted-window self-attention The role of this block is to learn local attention features at the scale determined by its position i in the stages sequence. It consists of n_i alternations of window multihead self-attention (W-MSA) and shifted-window multihead self-attention (SW-MSA), each followed by a 1-hidden layer MLP with expansion factor $\alpha = 4$ and GELU activation. In the W-MSA module, attention is limited to a window that contains $M \times M$ patches (see Fig. 2b). In the SW-MSA module, the same

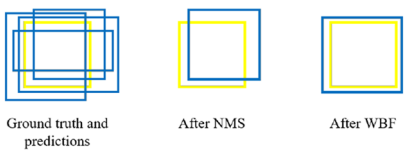

Fig. 3 Illustration of NMS and WBF outcomes for an ensemble of inaccurate predictions (blue: different models predictions; yellow: groundtruth)

windowing scheme is shifted by $(\frac{M}{2}, \frac{M}{2})$ to allow patches previously belonging to different windows to attend each other being now in the same window (see Fig. 2c). Layer normalization and skip connections are applied before and after each module, respectively, like in traditional vision transformers.

Patch merging The role of this block is to perform a downsampling-like operation similar to pooling in convolutional networks. To this end, groups of 2×2 neighboring patches are concatenated and forwarded to a linear layer that reduces the input dimensionality by a factor of 2. For example, in the first patch merging block, each patch is encoded into a vector of size C, thus the $4C$-dimensional concatenated patches are reduced to $2C$-dimensionality (see Fig. 2a). In the second patch merging block, the $8C$-dimensional concatenated patches are reduced to $4C$-dimensionality, and so on. Similar to convolutional networks, the the number of features increases together with the reduction along the spatial dimensions.

4.2.2 Architecture variants

We used the same architecture variants as the ones proposed in Liu et al. (2021b) to take advantage from their available ImageNet-pretrained models. The Swin base model (Swin-B) was built to have similar size and computation complexity as ViT-B. Three variants of the base model were introduced: Swin tiny (Swin-T), Swin small (Swin-S), and Swin large (Swin-L) which have around 0.25×, 0.5× and 2× the complexity and size of Swin-B, respectively. The architectures and their corresponding hyperparameters are provided in Table 2.

4.3 Object detection methods

In this study, we employ two object detection methods and then merge their predictions. The first, RepPoints (Yang et al. 2019), was used by the authors of the Swin Transformer achieving state-of-the-art object detection performance on COCO (Liu et al. 2021b). The second, Deformable Detection Transformer (DETR) (Zhu et al. 2020), is a novel transformer-based object detection model, which we found promising to combine with a transformer-based backbone. In the following, the two object detection heads are briefly illustrated.

4.3.1 RepPoints

RepPoints (Yang et al. 2019) is an anchor-free object detector which proposes a representation of objects as a set of sample points, suitable for both localization and recognition. The representative points learn to automatically organize themselves in a manner that bounds the spatial extent of an object and highlights semantically meaningful local areas when groundtruth localization and recognition targets are given for training. The training of RepPoints is driven jointly by object localization and recognition targets, such that the RepPoints are tightly bound by the groundtruth bounding box and guide the detector toward correct object classification.

4.3.2 Deformable DETR

DETR (Carion et al. 2020) is an end-to-end transformer-based object detection framework mainly characterized by the use of a set-based global loss which enforces unique predictions via bipartite matching and a transformer encoder–decoder architecture. While DETR removes the need of hand-designed components such as anchor generation, which directly encodes the prior knowledge of the task, it also suffers from limited feature spatial resolution and slow convergence. Deformable DETR (DDETR) (Zhu et al. 2020) aims to mitigate DETR issues by combining the best of deformable convolutions sparse spatial sampling and Transformers relation modeling capability. It also proposes a deformable attention module which attends to a restricted number of sample locations as a pre-filter for significant key components out of all the feature map pixels. DDETR replaces transformer attention modules processing feature maps by multiscale deformable attention modules.

4.4 Weighted boxes fusion

We employ weighted boxes fusion (WBF) (Solovyev et al. 2021) to merge predictions from different detectors. Unlike

Table 3 Hyperparameters of the best detection models, star

Model	Learning rate	Optimizer	Epochs	WBF weight
RetinaNet/ ResNet-50	7.81×10^{-5}	SGD	13	–
RepPoints/Swin-T	1.25×10^{-5}	AdamW	19	2.0
RepPoints/Swin-B	1.25×10^{-5}	AdamW	32	1.3
RepPoints/ ResNet-50	1.00×10^{-4}	SGD	22	1.7
RepPoints/ ResNet-101	1.00×10^{-4}	SGD	16	1.3
DDETR/Swin-T	1.25×10^{-5}	AdamW	28	1.7
DDETR/Swin-B	1.25×10^{-5}	AdamW	18	0.4
DDETR/ ResNet-50	1.25×10^{-5}	AdamW	24	0.4
DDETR/ ResNet-101	1.25×10^{-5}	AdamW	24	0.1

Non-Maximum Suppression (NMS) and soft-NMS methods that discard part of the predictions, WBF uses the confidence scores of all proposed bounding boxes to generate the averaged boxes (see Fig. 3) and is explicitly designed for ensembling boxes from different object detection models. We used a grid search to find the optimal weight for each detector.

5 Experiments

Our experiments are divided into four stages:

1. **Baseline mass detector**. We trained and tested a baseline mass detector that reproduces the work of Agarwal et al. (2020) which set the benchmark on the OMI-DB database, the same used in our study. To this end, we combined a ResNet-50 (He et al. 2016b) backbone and RetinaNet (Lin et al. 2017) object detection head with anchor boxes initialized as suggested in Agarwal et al. (2020).
2. **Mass detection with Swin Transformer backbones**. We trained and tested the two selected object detection models (see Sect. 4.3) with two variants of the Swin Transformer architecture: Swin-T and Swin-B. We discarded Swin-S since it exhibited very similar performances compared to the smaller and faster-to-train Swin-T throughout all experiments. In addition, we could not train Swin-L due to its large memory footprint exceeding the capability of our GPU.
3. **Mass detection with convolutional backbones**. We trained and tested the two selected object detection models with the convolutional backbones counterparts of Swin-T and Swin-B in terms of size and computational complexity, namely ResNet-50 and ResNet-101, respectively, as indicated in Liu et al. (2021b).
4. **Fusion of predictions**. We merged the predictions of mass detectors in three different scenarios: (i) detectors with Swin backbones only; (ii) detectors with convolutional backbones only; and (iii) all detectors. This was done to assess whether convolutional-based and transformer-based backbones could complement each other by extracting different features and provide an overall boosted performance.

5.1 Implementation

Training of all models was done on one NVIDIA Tesla V100 16 GB GPU. We used MMDetection (Chen et al. 2019), a PyTorch-based open source object detection toolset, to implement, train, and test all the architectures considered. MMDetection provides a collection of object detection models pretrained on the COCO dataset (Lin et al. 2014). Pretrained models of our selected object detection methods

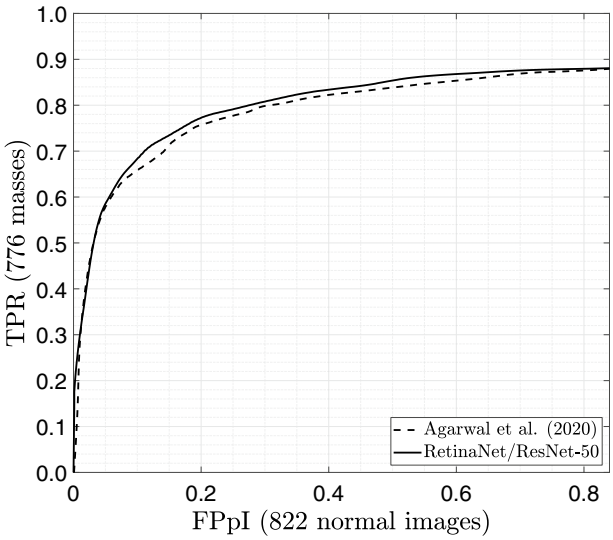

Fig. 4 Comparison between the FROC curve obtained by Agarwal et al. (2020) (taken from their paper) and our baseline detector

with a Swin Transformer backbone were not available, thus we used the pretrained ImageNet weights of Swin Transformer provided by the authors, along with our object detection methods pretrained with convolutional backbones for finetuning. Pretrained models of the object detection methods were available with convolutional backbones and were therefore used to finetune the convolutional models. In the view of the above, convolutional backbones had an advantage over transformer backbones since the the former were pretrained for object detection on COCO, whereas the latter were pretrained only for image classification on ImageNet.

5.2 Data preprocessing

All images were resized such that the height and width were at most 800 and 1333 pixels, respectively, while keeping the original aspect ratio. Due to memory limitations, in the case of DDETR the maximum height and width were 600 and 1000 pixels, respectively. In addition, pixel values were normalized to zero mean and unit standard deviation.

5.3 Data augmentation

During training, we applied with probability 50% one of the following data augmentation techniques to each incoming image: (i) horizontal flip; (ii) random crop; (iii) contrast transformation, with magnitude values of [0.4, 0.8, 1.5]; and (iv) brightness transformation, with magnitude values of [0.3, 0.7, 1.3]. For deeper backbones (Swin-B and Resnet 101), the probabilities were increased to 60%. For the baseline RetinaNet-based detector, we applied only horizontal

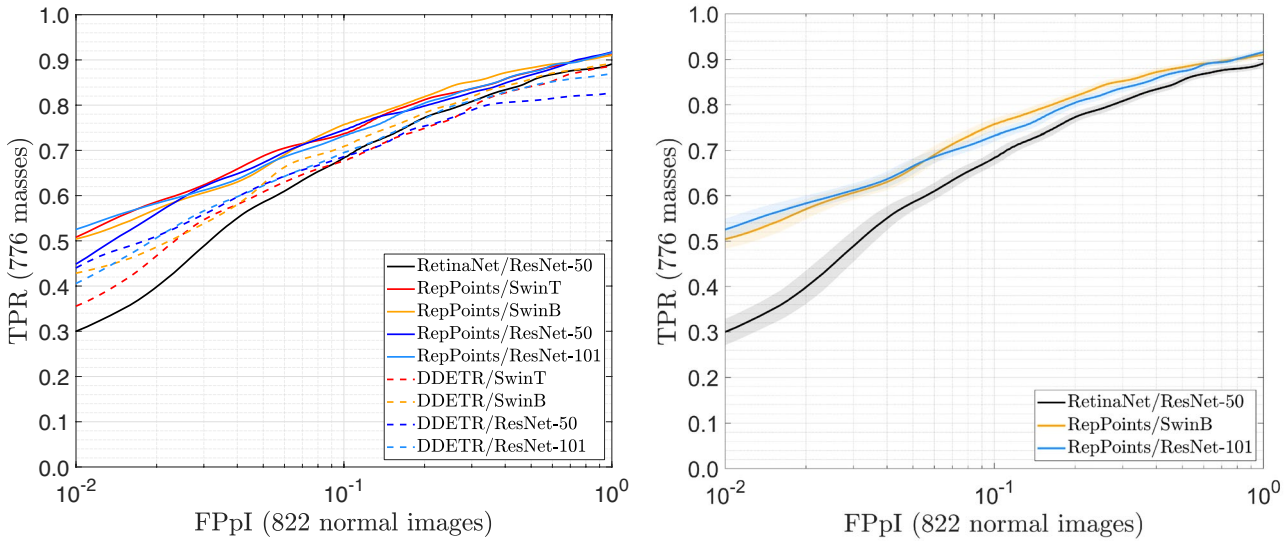

Fig. 5 Average FROC curves obtained from 10,000 bootstrap iterations comparing all detectors (left) and baseline, best convolutional-based, and best transformer-based detectors (right). Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPR axis

Table 4 Performance results of single models averaged from 10,000 bootstrap iterations

Model	AUC	TPR
RetinaNet/ResNet-50	81.0%	68.3%
RepPoints/Swin-T	84.5%	73.7%
RepPoints/Swin-B	**85.0%**	**75.7%**
RepPoints/ResNet-50	84.0%	74.5%
RepPoints/ResNet-101	84.3%	73.2%
DDETR/Swin-T	80.4%	67.7%
DDETR/Swin-B	82.1%	70.8%
DDETR/ResNet-50	77.8%	68.6%
DDETR/ResNet-101	80.6%	69.5%

Highest performance for each column is marked in bold

flip as suggested in the reference work of Agarwal et al. (2020).

5.4 Training hyperparameters

All the models were trained for a maximum of 100 epochs in batches of 2 images using the backpropagation algorithm, two different optimizers (Stochastic Gradient Descent and AdamW (Loshchilov and Hutter 2017)) and learning rates in the range $[10^{-3}, 10^{-6}]$. The best model was selected as the one achieving the highest mean Average Precision (mAP) over IoU thresholds from 0.1 to 0.5 (step 0.05). This metric was also used to monitor the performance and for early stopping, which occurred in all experiments between epoch 13 and epoch 32. Each selected model was then assigned a weight in the WBF stage by performing a grid search on weights between 0.1 and 2 with a step of 0.3. Table 3 presents the details of the best models selected. All the optimizations were carried out on the validation set.

5.5 Performance evaluation

To assess the performance of the compared methods, we calculated lesion-based free receiver operating characteristic (FROC) curves that report the True Positive Rate (TPR) of the detected lesions versus the average number of False Positives per Image (FPpI) by varying the decision threshold applied to the scores associated to the detected boxes. A predicted box was considered a true positive when its IoU with the groundtruth mass was equal or greater than 10% following the criterion used in Agarwal et al. (2020). All predictions on normal images were counted as false positives.

From the FROC curves, two performance measures were extracted: (i) the area under the curve (AUC) in the FPpI range [0, 1] that is a widely adopted range for evaluating CADe systems and it is also used in Agarwal et al. (2020); and (ii) the TPR at FPpI=0.1 that can be a clinically useful operating point considering that radiologists' specificity ranges from 95 to 98% (Salim et al. 2020) which is orders of magnitude better than operating points near FPpI=1 commonly evaluated in the mass detection literature.

5.6 Statistical analysis

The bootstrap method (Samuelson and Petrick 2006) was applied to test the statistical significance of differences in

AUC and TPR between the compared methods, as typically done when comparing CAD systems performances (Ma et al. 2013; Hupse and Karssemeijer 2009; Bria et al. 2014; Mordang et al. 2016; Kooi et al. 2017; Wang and Yang 2019). Patients were sampled with replacement 10,000 times, with each bootstrap containing the same number of patients as the original set. At each bootstrapping iteration, FROC curves were recalculated for each method, and ΔAUC and ΔTPR

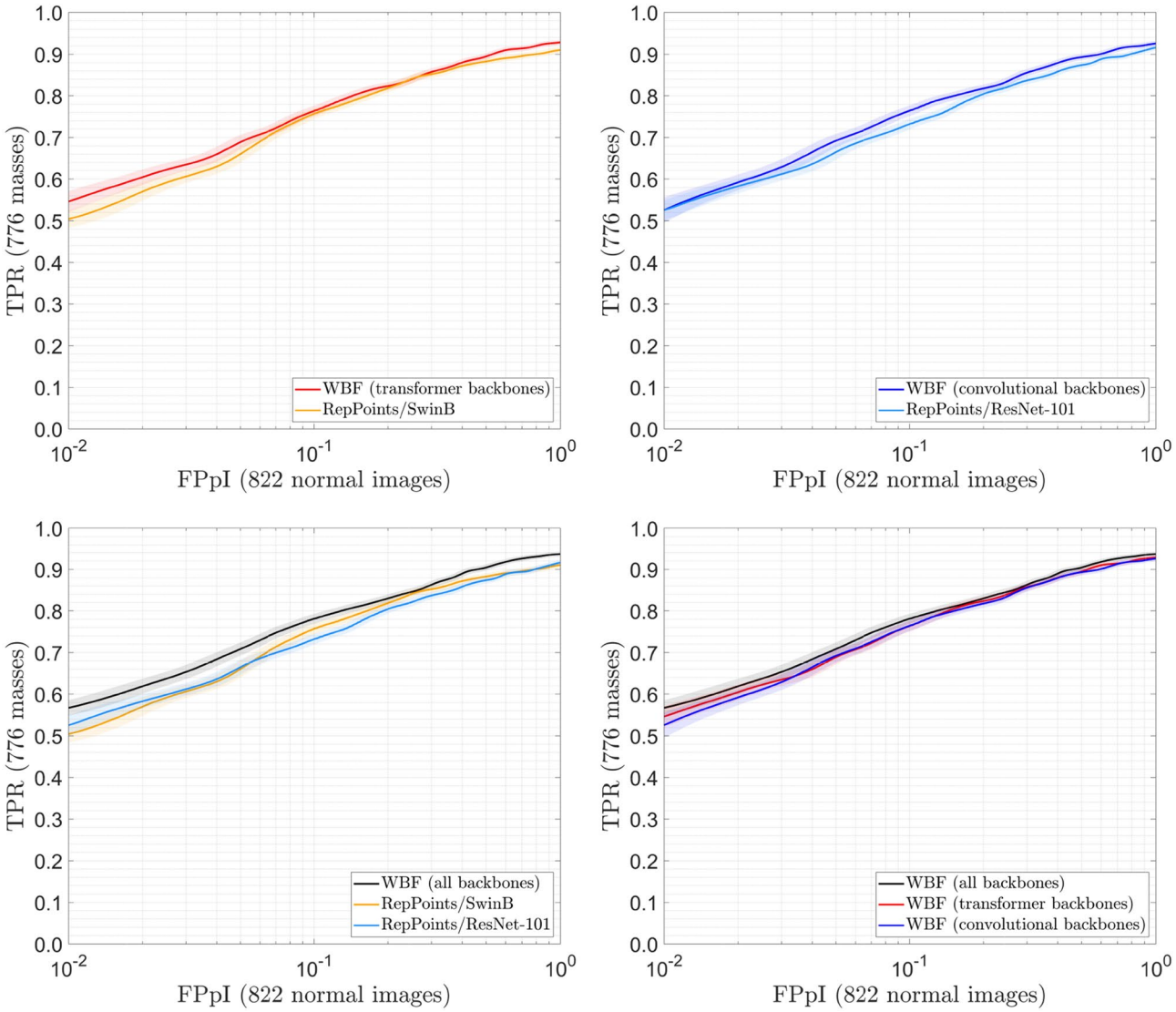

Fig. 6 Average FROC curves obtained from 10,000 bootstrap iterations comparing transformer fusion with best single transformer model (top-left), convolutional fusion with best single convolutional model (top-right), transformer+convolutional fusion with best single transformer and convolutional models (bottom-left), and all fusion models (bottom-right). Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPR axis

Table 5 Performance results of fusion models averaged from 10,000 bootstrap iterations

WBF	Models fused	AUC	TPR
Transformer backbones	RepPoints/Swin-T RepPoints/Swin-B DDETR/Swin-T DDETR/Swin-B	86.4%	76.3%
Convolutional backbones	RepPoints/ResNet-50 RepPoints/ResNet-101 DDETR/ResNet-50 DDETR/ResNet-101	86.1%	76.4%
All backbones	RepPoints/Swin-T RepPoints/Swin-B DDETR/Swin-T DDETR/Swin-B RepPoints/ ResNet-50 RepPoints/ResNet-101 DDETR/ResNet-50 DDETR/ResNet-101	**87.4**%	**78.1**%

Highest performance for each column is marked in bold

Table 6 Statistical comparisons in terms of AUC and TPR differences obtained from 10,000 boostrap iterations

Description	Model	Compared to	ΔAUC	ΔTPR
Transformer backbones vs. convolutional backbones	RepPoints/Swin-T	RepPoints/ResNet-50	+0.5%($p = 0.2789$)	−0.8%($p = 0.6948$)
	RepPoints/Swin-B	RepPoints/ResNet-101	+0.7%($p = 0.1985$)	+2.5%($p = 0.0869$)
	DDETR/Swin-T	DDETR/ResNet-50	**+2.6**%($p = 0.0106$)	−0.9%($p = 0.6907$)
	DDETR/Swin-B	DDETR/ResNet-101	+1.5%($p = 0.0608$)	+1.4%($p = 0.2464$)
Deeper backbones vs. less deep backbones	RepPoints/Swin-B	RepPoints/Swin-T	+0.5%($p = 0.2439$)	+2.0%($p = 0.1261$)
	RepPoints/ResNet-101	RepPoints/ResNet-50	+0.3%($p < 0.3571$)	−1.3%($p = 0.7969$)
	DDETR/Swin-B	DDETR/Swin-T	**+1.8**%($p = 0.0220$)	**+3.1**%($p = 0.0360$)
	DDETR/ResNet-101	DDETR/ResNet-50	**+2.8**%($p = 0.0013$)	+0.9%($p = 0.2910$)
RepPoints vs. DDETR	RepPoints/Swin-B	DDETR/Swin-B	**+2.8**%($p < 0.0018$)	**+4.8**%(**p < 0.0059**)
	RepPoints/ResNet-101	DDETR/ResNet-101	**+3.6**%($p < 0.0001$)	**+3.7**%(**p = 0.0199**)
Our best single models vs. state-of-the-art	RepPoints/Swin-B	RetinaNet/ResNet-50	**+4.0**%($p < 0.0001$)	**+7.4**%($p < 0.0001$)
	RepPoints/ResNet-101	RetinaNet/ResNet-50	**+3.3**%($p < 0.0001$)	**+4.9**%($p = 0.0118$)
	RepPoints/ResNet-50	RetinaNet/ResNet-50	**+3.0**%($p < 0.0001$)	**+6.2**%($p = 0.0011$)
WBF (transformer) vs. WBF (convolutional)	WBF (transformer)	WBF (convolutional)	+0.3%($p = 0.2914$)	0.0%($p = 0.5257$)
WBF (single-kind) vs. best singlemodel	WBF (convolutional)	RepPoints/ResNet-101	**+1.8**%($p < 0.0001$)	**+3.2**%($p = 0.0064$)
	WBF (transformer)	RepPoints/Swin-B	**+1.4**%($p = 0.0088$)	+0.7%($p = 0.3442$)
WBF (all) vs. best single models	WBF (all)	RepPoints/Swin-B	**+2.5**%($p < 0.0001$)	**+2.4**%($p = 0.0421$)
	WBF (all)	RepPoints/ResNet-101	**+3.2**%($p < 0.0001$)	**+4.9**%($p = 0.0007$)
WBF (all) vs. WBF (single-kind)	WBF (all)	WBF (convolutional)	**+1.4**%($p = 0.0010$)	+1.7%($p = 0.0984$)
	WBF (all)	WBF (transformer)	**+1.0**%($p < 0.0001$)	**+1.8**%($p = 0.0224$)
Our best WBF model vs. state-of-the-art	WBF (all)	RetinaNet/ResNet-50	**+6.5**%($p < 0.0001$)	**+9.8**%($p < 0.0001$)

Statistically significant differences (p-value < 0.05) are listed in bold

were evaluated for each pair of methods. p-values were computed as the fraction of evaluated metrics populations that were negative or zero, corresponding to cases where the first method did not outperform the second method under comparison (null hypothesis). Performance differences were considered statistically significant if $p < 0.05$.

6 Results and discussion

6.1 Baseline

In Fig. 4 the FROC curve originally published in the paper of Agarwal et al. (2020) and the FROC curve of our baseline mass detector are reported. It can be seen that the two curves are almost identical. Furthermore, Agarwal et al. (2020) reported TPR=0.87 at 0.84 FPpI while our model reaches TPR=0.88 at 0.84 FPpI. These results indicate that we were able to fully replicate the method and the results of Agarwal et al. (2020), thus it can be used as baseline for comparison with our proposed single and fusion models.

6.2 Single models

In Fig. 5 and Table 4 are reported the FROC curves and performance results, respectively, of all experimented detectors to compare transformer-based and convolutional-based backbones. Statistical comparisons are reported in Table 6.

Comparison between transformer and convolutional backbones On average, transformer backbones provided +1.3% AUC and +0.5% TPR improvements over their convolutional counterparts. The best transformer-based model was RepPoints/Swin-B achieving 85.0% AUC and 75.7% TPR, with an improvement of +0.7% AUC ($p = 0.1985$) and +2.5% TPR ($p = 0.0869$) over the best convolutional-based model RepPoints/ResNet-101. This suggests that transformer backbones are a viable and promising alternative to convolutional backbones for mass detection. We believe that transformers have further potential to exploit, considering that: (i) convolutional backbones benefited from pretrained convolutional-based object detection heads whereas transformer backbones were only pretrained for image classification; (ii) we used the Swin Transformer 'as is' without modifying important architecture hyperparameters such as patch size, window size, and number of layers since that would have required training from scratch without pretrained weights, which in turns would have required a large amount of data; and (iii) an improved second version of Swin capable of handling higher resolution images has been just released (Liu et al. 2021a) along with other general-purpose transformer backbones like Nested

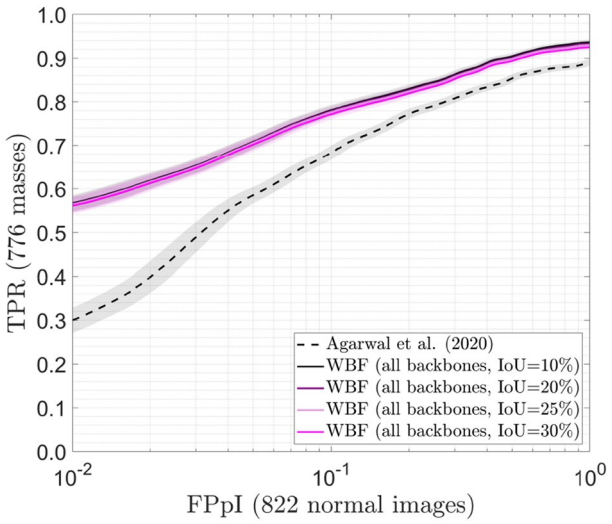

Fig. 7 FROC curves of the proposed method (WBF all backbones) with different IoU thresholds

Table 7 Computational complexity of all experimented models

Model	TTpE	#param	FLOPs
RetinaNet/ResNet-50	14.6	36 M	206 G
RepPoints/Swin-T	10.6	37 M	195 G
RepPoints/Swin-B	17.3	97 M	428 G
RepPoints/ResNet-50	7.8	37 M	190 G
RepPoints/ResNet-101	10.1	56 M	266 G
DDETR/Swin-T	18.8	39 M	516 G
DDETR/Swin-B	23.7	98 M	749 G
DDETR/ResNet-50	10.2	40 M	195 G
DDETR/ResNet-101	11.8	59 M	271 G
WBF (convolutional)	39.9	192 M	922 G
WBF (transformer)	70.4	271 M	1,888 G
WBF (all)	110.3	463 M	2,810 G

TTpE Training Time per Epoch (in min), *FLOPs* FLoating point Operations Per Second

Hierarchical Transformer (NeST) (Zhang et al. 2022) and Pyramid Vision Transformer (PVT) (Wang et al. 2021).

Comparison between different depths Deeper backbones Swin-B and ResNet-101 yielded an average improvement of +1.4% AUC and +1.2% TPR over less deep backbones Swin-T and ResNet-50. This was an expected result, since deeper backbones are better suited to extract multiscale features from high resolution images like mammograms. The difference is higher when comparing Swin-T and Swin-B, the latter outperforming the former on average by +1.2% AUC and +2.6% TPR. We believe this is due to the different embedding size ($C = 96$ for Swin-T and $C = 128$ for Swin-B) which in turns leads to an increased number of attention heads for Swin-B (4 instead of 3 for each layer) and thus of features extracted. This makes promising trying with Swin-L which has $C = 192$ and 6 heads per layer, however we could not train such a big model because of GPU memory limitations.

Comparison between different detection heads RepPoints detectors statistically significantly outperformed their DDETR counterparts, with an average improvement of +4.2% AUC and +5.1% TPR. This was due to the different resolutions of the input images (see Sect. 5.4), indicating that despite masses are orders-of-magnitude bigger than other mammographic lesions like calcifications that require full-resolution images to be detected, the impact of image downsampling on the overall mass detection performance is still not negligible. Another interesting result is the improvement in AUC of transformer over convolutional backbones with DDETR (+2.0%) compared with the improvement with RepPoints (+0.6%), nearly three times smaller. This suggests that DDETR, being a transformer-based head, can benefit from transformer-extracted features more than RepPoints, which is a convolutional-based head. Thus the potential of DDETR combined with a transformer backbone could be fully disclosed when training with higher resolution.

Comparison with state-of-the-art mass detection Both transformer and convolutional backbones combined with RepPoints statistically significantly outperformed the baseline reference method reproduced from the work of Agarwal et al. (2020). This can be appreciated qualitatively from Fig. 5 (right) and quantitatively from Tables 4, 5 and 6 with an average increment of +3.5% AUC and +5.9% TPR over the baseline and a massive improvement of +4.0% AUC ($p < 0.0001$) and +7.4% TPR ($p < 0.0001$) yielded by the best model RepPoints/Swin-B. However, it must be noted that most of the improvement is provided by the RepPoints head, since with the same convolutional backbone (ResNet-50) RepPoints/ResNet-50 statistically significantly surpasses the baseline RetinaNet/ResNet-50 by +3.0% AUC and +6.2% TPR. This suggests that object detectors based on point set representations, like RepPoints, are well suited for mass detection since they can account for the shape and positions of semantically important local areas of the lesions, whereas detectors like RetinaNet consider equally all subregions within an anchor box impeding finer feature extraction.

6.3 Fusion models

In Fig. 6 and Table 5 are reported the FROC curves and performance results, respectively, of all fusion models. Statistical comparisons are reported in Table 6.

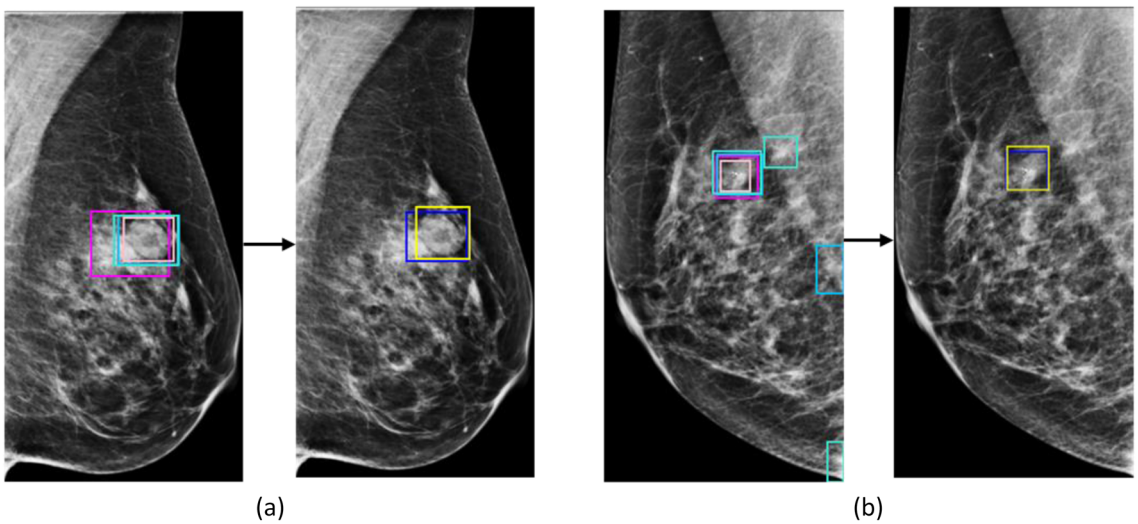

Fig. 8 WBF fusion of convolutional backbones (**a**) and transformer backbones (**b**). Groundtruth bounding box and fusion box are displayed in blue and yellow colors, respectively

Fusion of single-kind models WBF applied to the four detectors based on transformer backbones yielded an improvement of +1.4% AUC ($p = 0.0088$) and +0.7% TPR ($p = 0.3442$) over the best single transformer model RepPoints/Swin-B. Similarly, WBF applied to the four detectors based on convolutional backbones yielded an improvement of +1.8% AUC ($p < 0.0001$) and +3.2% TPR ($p = 0.0064$) over the best single convolutional model RepPoints/ResNet-101. The performances of the two fusion models were statistically significantly similar, with a negligible advantage of the transformer fusion over the convolutional fusion (+0.3% AUC). In both cases, the positive impact of WBF can be qualitatively appreciated on the examples presented in Fig. 8, where it gives more accurate prediction coordinates and also discards wrong predictions.

Fusion of all models WBF applied to all eight detectors, both transformer- and convolutional-based, achieved an AUC of 87.4% and a TPR of 78.1% statistically significantly surpassing the best single model RepPoints/Swin-B by +2.5% AUC and +2.4% TPR and the best single-kind fusion model that merged all transformer backbones by +1.0% AUC and +1.8% TPR. This result suggests that convolutional and transformer backbones extract different features that complement each other when combined together. It also can be observed that the WBF weights (see Table 3, last column) assigned more importance to transformer backbones, with an aggregate of 5.4 compared to 3.5 of the convolutional backbones, suggesting that the former had a more positive impact when performing the fusion of the predictions.

Robustness to higher IoU We carried out additional experiments to assess the performance of WBF applied to all models, while modifying the IoU threshold used to match a predicted box with a groundtruth mass (see Sect. 5.5). Previously we used IoU=0.1, whereas on this set of experiments we increase the threshold to 0.2, 0.25 and 0.3. The obtained FROC curves are plotted in Fig. 7, from which it can be observed that these models exhibit similar performances, since there was no statistically significant difference in AUC and TPR. At the same time, all of them still largely outperform the baseline model, which instead was evaluated with the more favorable IoU=0.1.

6.4 Computational complexity

We provide in Table 7 the computational complexity of all experimented models in terms of training time per epoch, number of learnable parameters, and floating point operations per second. Overall, transformer-based models exhibited a nearly double computational complexity compared to convolutional-based models. This was mainly due by the contribution of the backbone Swin-B, which in Liu et al. (2021b) was reported to have more than triple computational complexity compared to the more lightweight backbone Swin-T.

Further, it was in general more difficult to train transformer models compared to convolutional models due to higher memory requirements with same input dimensions, time to convergence, and sensitivity to the chosen optimizer and learning rate configuration.

7 Conclusions

This study focuses on the detection of masses on digital mammograms using transformer-based architectures on the large-scale publicly available dataset OMI-DB. To our knowledge, the proposed work is the first to attempt implementing a transformer backbone for mass detection in mammograms, resulting in models that outperform previous state-of-the-art methods. It was shown that transformer-based models can be pretrained on natural images and be successfully finetuned to detect masses in mammograms. The implemented models achieved promising results on this task and showed superior performances to their convolutional counterparts. Compared to the state-of-the-art model previously proposed and tested on OMI-DB, our mass detection models achieved statistically significantly higher area under the FROC and sensitivity, up to 4.0% AUC and 7.4% TPR improvement obtained with a transformer backbone. Additionally, combining the predictions of both convolutional and transformer models using weighted boxes fusion results in a massive improvement of 6.5% AUC and 9.8% TPR. These results well demonstrate the potential of transformer backbones in detection tasks on medical images, thus our future work will address the implementation and adaptation of other transformer-based architectures that are quickly arising in the computer vision literature.

Acknowledgements This work was supported by MIUR (Minister for Education, University and Research, Law 232/216, Department of Excellence). Amparo S. Betancourt T. holds an EACEA Erasmus+ grant for the master in Medical Imaging and Applications (MAIA).

Data availibility The OMI-DB dataset (Halling-Brown et al. 2020) employed in the current study is publicly available at https://medphys.royalsurrey.nhs.uk/omidb/. The list of images of the OMI-H-MD subset that we extracted and used in our experiments are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agarwal R, Díaz O, Yap MH, Llado X, Marti R (2020) Deep learning for mass detection in Full Field Digital Mammograms. Comput Biol Med 121:103774

Aly GH, Marey M, El-Sayed SA, Tolba MF (2021) YOLO based breast masses detection and classification in Full-Field Digital Mammograms. Comput Methods Programs Biomed 200:105823

Balleyguier C, Kinkel K, Fermanian J, Malan S, Djen G, Taourel P, Helenon O (2005) Computer-aided detection (CAD) in mammography: does it help the junior or the senior radiologist? Eur J Radiol 54(1):90–96

Bria A, Karssemeijer N, Tortorella F (2014) Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications. Med Image Anal 18(2):241–252

Cao H, Pu S, Tan W, Tong J (2021) Breast mass detection in digital mammography based on anchor-free architecture. Comput Methods Programs Biomed 205:106033

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer

CDC (2022) Breast cancer screening guidelines for women. https://www.cdc.gov/cancer/breast/pdf/breast-cancer-screening-guidelines-508.pdf. Accessed: 2022-05-20

Chen X, Zhang K, Abdoli N, Gilley PW, Wang X, Liu H, Zheng B, Qiu Y (2022) Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. Diagnostics 12(7):1549

Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J et al (2019) MMDetection: open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. Ieee

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Ferrari RJ, Rangayyan RM, Desautels JL, Borges R, Frere AF (2004) Automatic identification of the pectoral muscle in mammograms. IEEE Trans Med Imaging 23(2):232–245

Halling-Brown MD, Warren LM, Ward D, Lewis E, Mackenzie A, Wallis MG, Wilkinson LS, Given-Wilson RM, McAvinchey R, Young KC (2020) OPTIMAM Mammography image database: a large-scale resource of mammography images and clinical data. Radiology 3(1):e200103

Heath MD, Bowyer KW (2000) Mass detection by relative image intensity. In: Proceedings of the 5th International Workshop on Digital Mammography (IWDM-2000), pp 219–225

He K, Zhang X, Ren S, Sun J (2016a) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

He K, Zhang X, Ren S, Sun J (2016b) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708

Hupse R, Karssemeijer N (2009) Use of normal tissue context in computer-aided detection of masses in mammograms. IEEE Trans Med Imaging 28(12):2033–2041

Johnson KB, Wei W-Q, Weeraratne D, Frisse ME, Misulis K, Rhee K, Zhao J, Snowdon JL (2021) Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 14(1):86–93

Kamran SA, Hossain KF, Tavakkoli A, Bebis G, Baker S (2022) Swinsftnet: spatial feature expansion and aggregation using swin transformer for whole breast micro-mass segmentation. arXiv preprint arXiv:2211.08717

Ke L, Mu N, Kang Y (2010) Mass computer-aided diagnosis method in mammogram based on texture features. In: 2010 3rd International Conference on Biomedical Engineering and Informatics, Volume 1, . 354–357. IEEE

Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N (2017) Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 35:303–312

Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

Kwok SM, Chandrasekhar R, Attikiouzel Y, Rickard MT (2004) Automatic pectoral muscle segmentation on mediolateral oblique view mammograms. IEEE Trans Med Imaging 23(9):1129–1140

Lbachir IA, Daoudi I, Tallal S (2021) Automatic computer-aided diagnosis system for mass detection and classification in mammography. Multimed Tools Appl 80(6):9493–9525

Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988

Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision, pp 740–755. Springer

Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, Ning J, Cao Y, Zhang Z, Dong L et al (2021) Swin transformer V2: scaling up capacity and resolution. arXiv preprint arXiv:2111.09883

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10012–10022

Li F, Zhang H, Liu S, Zhang L, Ni LM, Shum H-Y et al (2022) Mask dino: towards a unified transformer-based framework for object detection and segmentation. arXiv preprint arXiv:2206.02777

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101

Ma H, Bandos AI, Rockette HE, Gur D (2013) On use of partial area under the roc curve for evaluation of diagnostic performance. Stat Med 32(20):3449–3458

Malliori A, Pallikarakis N (2022) Breast cancer detection using machine learning in digital mammography and breast tomosynthesis: a systematic review. Health Technol 1–18

Méndez AJ, Tahoces PG, Lado MJ, Souto M, Correa J, Vidal JJ (1996) Automatic detection of breast border and nipple in digital mammograms. Comput Methods Programs Biomed 49(3):253–262

Molinara M, Marrocco C, Tortorella F (2013) Automatic segmentation of the pectoral muscle in mediolateral oblique mammograms. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pp 506–509. IEEE

Mordang J-J, Janssen T, Bria A, Kooi T, Gubern-Mérida A, Karssemeijer N (2016) Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In: International Workshop on Digital Mammography, pp 35–42. Springer

Mughal B, Sharif M, Muhammad N (2017) Bi-model processing for early detection of breast tumor in CAD system. Eur Phys J Plus 132(6):1–14

Patel BC, Sinha G, Soni D (2019) Detection of masses in mammographic breast cancer images using modified histogram based adaptive thresholding (MHAT) method. Int J Biomed Eng Technol 29(2):134–154

Petrick N, Chan H-P, Sahiner B, Wei D (1996) An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. IEEE Trans Med Imaging 15(1):59–67

Petrick N, Chan H-P, Sahiner B, Helvie MA (1999) Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. Med Phys 26(8):1642–1654

Punitha S, Amuthan A, Joseph KS (2018) Benign and malignant breast cancer segmentation using optimized region growing technique. Future Comput Inform J 3(2):348–358

Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) Ai in health and medicine. Nat Med 28(1):31–38

Redmon J, Divvala S, Girshick R, Farhadi A (2016a) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788

Redmon J, Divvala S, Girshick R, Farhadi A (2016b) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788

Ren S, He K, Girshick R, Sun J(2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28

Ribli D, Horváth A, Unger Z, Pollner P, Csabai I (2018) Detecting and classifying lesions in mammograms with deep learning. Sci Rep 8(1):1–7

Salim M, Dembrower K, Eklund M, Lindholm P, Strand F (2020) Range of radiologist performance in a population-based screening cohort of 1 million digital mammography examinations. Radiology 297(1):33–39

Samuelson FW, Petrick N (2006) Comparing image detection algorithms using resampling. In: IEEE Int. Symp. Biomed. Imag., pp 1312–1315

Sankatsing VD, van Ravesteyn NT, Heijnsdijk EA, Looman CW, van Luijt PA, Fracheboud J, den Heeten GJ, Broeders MJ, de Koning HJ (2017) The effect of population-based mammography screening in Dutch municipalities on breast cancer mortality: 20 years of follow-up. Int J Cancer 141(4):671–677

Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H (2022) Transformers in medical imaging: a survey. arXiv preprint arXiv:2201.09873

Solovyev R, Wang W, Gabruseva T (2021) Weighted boxes fusion: Ensembling boxes from different object detection models. Image Vis Comput 107:104117

Su Y, Liu Q, Xie W, Hu P (2022) Yolo-logo: a transformer-based yolo segmentation model for breast mass detection and segmentation in digital mammograms. Comput Methods Programs Biomed 106903

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Cancer J Clin 71(3):209–249

Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312

Te Brake GM, Karssemeijer N (1999) Single and multiscale detection of masses in digital mammograms. IEEE Trans Med Imaging 18(7):628–639

Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM (2021) Medical transformer: gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 36–46. Springer

Wang Z, Yu G, Kang Y, Zhao Y, Qu Q (2014) Breast tumor detection in digital mammography based on extreme learning machine. Neurocomputing 128:175–184

Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L (2021) Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 568–578

Wang J, Yang Y (2019) A hierarchical learning approach for detection of clustered microcalcifications in mammograms. In : 2019 IEEE International Conference on Image Processing (ICIP), pp 804–808

Wei Y, Hu H, Xie Z, Zhang Z, Cao Y, Bao J, Chen D, Guo B (2022) Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. arXiv preprint arXiv:2205.14141

Yan Y, Conze P-H, Lamard M, Quellec G, Cochener B, Coatrieux G (2021) Towards improved breast mass detection using dual-view mammogram matching. Med Image Anal 71:102083

Yang Z, Liu S, Hu H, Wang L, Lin S(2019) Reppoints: point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9657–9666

Yu X, Wang S-H, Zhang Y-D(2022) Multiple-level thresholding for breast mass detection. J King Saud Univ-Comput Inf Sci

Zhang L, Li Y, Chen H, Wu W, Chen K, Wang S (2022) Anchor-free yolov3 for mass detection in mammogram. Expert Syst Appl 191:116273

Zhang Z, Zhang H, Zhao L, Chen T, Arik SÖ, Pfister T (2022) Nested hierarchical transformer: towards accurate, data-efficient and interpretable visual understanding. Proc AAAI Conf Artif Intell 36:3417–3425

Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, Ni LM, Shum H-Y (2022) Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605

Zhou Z, Shin J, Zhang L, Gurudu S, Gotway M, Liang J (2017) Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7340–7351

Zhu C, He Y, Savvides M (2019) Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 840–849

Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159