Computer Methods and Programs in Biomedicine 221 (2022) 106903

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms

Yongye Su[a], Qian Liu[a,b], Wentao Xie[a], Pingzhao Hu[a,b,c,*]

[a] Department of Biochemistry and Medical Genetics, University of Manitoba, Room 308-Basic Medical Sciences Building, 745 Bannatyne Avenue, Winnipeg, Manitoba R3E 0J9, Canada
[b] Department of Computer Science, University of Manitoba, Winnipeg, Canada
[c] CancerCare Manitoba Research Institute, CancerCare Manitoba, Winnipeg, Canada

ARTICLE INFO

Article history:
Received 12 December 2021
Revised 15 May 2022
Accepted 22 May 2022

Keywords:
Breast cancer
Mass detection
Mass segmentation
Deep learning
Transformer

ABSTRACT

Background and objective: Both mass detection and segmentation in digital mammograms play a crucial role in early breast cancer detection and treatment. Furthermore, clinical experience has shown that they are the upstream tasks of pathological classification of breast lesions. Recent advancements in deep learning have made the analyses faster and more accurate. This study aims to develop a deep learning model architecture for breast cancer mass detection and segmentation using the mammography.
Methods: In this work we proposed a double shot model for mass detection and segmentation simultaneously using a combination of YOLO (You Only Look Once) and LOGO (Local-Global) architectures. Firstly, we adopted YoloV5L6, the state-of-the-art object detection model, to position and crop the breast mass in mammograms with a high resolution; Secondly, to balance training efficiency and segmentation performance, we modified the LOGO training strategy to train the whole images and cropped images on the global and local transformer branches separately. The two branches were then merged to form the final segmentation decision.
Results: The proposed YOLO-LOGO model was tested on two independent mammography datasets (CBIS-DDSM and INBreast). The proposed model performs significantly better than previous works. It achieves true positive rate 95.7% and mean average precision 65.0% for mass detection on CBIS-DDSM dataset. Its performance for mass segmentation on CBIS-DDSM dataset is F1-score=74.5% and IoU=64.0%. The similar performance trend is observed in another independent dataset INBreast as well.
Conclusions: The proposed model has a higher efficiency and better performance, reduces computational requirements, and improves the versatility and accuracy of computer-aided breast cancer diagnosis. Hence it has the potential to enable more assistance for doctors in early breast cancer detection and treatment, thereby reducing mortality.

1. Introduction

Breast cancer is the most frequently diagnosed cancer and one of the leading common cancer-caused death for females [1,2]). One of the most commonly used early breast cancer screen methods is mammography [3]. Currently, the interpretation of mammography is still done manually by experienced radiologists. When the density of breast tissue is too high or the lesion is too small, it is easy to have a false negative observation [4]. Therefore, providing the radiologist with powerful computer-aided diagnosis (CAD) tools is one approach to improve mammographic interpretation and decision-making especially when the lesions are easy to be missed by manual detection [5]. Specifically, an automatic, accurate and efficient mass detection and segmentation model could be helpful for both manual diagnosis and automatic masses classification of breast cancer.

With the advancement of recent developments in deep learning, more and more new generation CAD models are constantly being created to explain mammography. In the past a few years, most deep learning- based image segmentation models use deep convolutional neural networks (CNNs) based U-Net (Olaf [6]) architecture to achieve their best performance, such as U-Net++ [7], Attention U-Net [8], DenseUNet [9], R2U-Net [10], UNet 3+ [11], Connected-Unet [12] and others [13–15]. CNNs could automatically generate important features from data in different domains with

* Corresponding author at: Department of Biochemistry and Medical Genetics, University of Manitoba, Room 308-Basic Medical Sciences Building, 745 Bannatyne Avenue, Winnipeg, Manitoba R3E 0J9, Canada.
E-mail address: pingzhao.hu@umanitoba.ca (P. Hu).

https://doi.org/10.1016/j.cmpb.2022.106903
0169-2607/

different abstract levels. This advantage makes it the most popular technique in deep learning-based computer vision research for the past few years. Currently, some of the high configuration GPUs are designed for the convolution calculation, which makes the CNNs faster and eventually allows CNN to continuously flourish in the field of computer vision. However, according to Valanarasu et al. [16], one major limitation of these CNN models is that they pay too much attention to local square of pixels. As a matter of the fact, a newly emerged deep learning architecture, named Vision Transformer (ViT), has recently been leading its trend for replacing CNNs in computer vision [17]. Comparing with traditional CNNs, ViT is powerful to capture both local and global or long-range visual dependencies through its self-attention mechanisms (either course-gained global attention or fine-grained local attention). Furthermore, the attention mechanism of the ViT could increase its interpretability, thus reducing the fear of "Black-box" when applying it in critical areas such as healthcare field [18,19]. These advantages of ViT make it a good fit to solve the mammography segmentation problem. The main challenge of ViT-based medical image segmentation model is that training such a model needs strong computing power [20]. Due to the limitations of GPU and other hardware conditions, current computing resources might only segment low-resolution images. Therefore, to some extent, this reminds us to develop a more accurate and computing efficient automatic segmentation system. Recently, there are works that combined the advantages of both the ViT and CNNs. For example, CvT introduced convolutions into ViT to achieve improved performance and efficiency, but it was not designed for solving segmentation problem [21]. TransUNet is the first transformer-based medical image segmentation framework that outperformed other state-of-the-art segmentation models [22]. It is also a combination of CNNs and ViT. Medical Transformer (MedT) proposed a gated axial transformer model structure which consists of convolutional layers and gated axial-attention layers [16]. The gated axial transformer blocks are then equipped with a LOGO (Local-Global) training strategy, which could fully utilize the medical image data to tackle with the small sample size issue [16].

Breast lumps usually only appear in a small area on a mammogram, which is called region of interesting (ROI). Yan et al. proposed that if the ROI could be detected first, it will benefit the segmentation task later [23]. They thus used a YoloV3 [24] as detector to identify the ROI from the mammograms, then passed the detected ROI into a CNN called V19U-net++ [7] for further segmentation. However, the ROI detection model YoloV3 they used has been updated to YoloV5L6 [25,26] now. They also made YoloV3 to detect only one breast mass ROI per image. However, in practice, it is possible to have more than one mass in a breast mammography image. Furthermore, their segmentation model V19U-net++ is a traditional CNN model, which could be improved by adapting transformer into its architecture.

Inspired by the MedT [16] and Jocher et al.'s two-stage detection-segmentation workflow, we proposed a novel breast mass segmentation model that first uses YoloV5 [25] and its latest model [26] to detect breast mass ROI in a mammogram, crop it directly from the high-resolution image, and finally use an updated LOGO training strategy [27,28] to segment mass from the cropped images. Our contributions are mainly in two aspects:

(1) We first tested the state-of-the-art object detection model, YoloV5, to detect breast mass in mammograms, and provided the specific cropped local images for later segmentation analysis.
(2) The advantage of ViT in considering long-range dependencies was fully utilized in our model. The LOGO structure of our model not only greatly improved the segmentation resolution at the original pixel level, but also maintained the positional accuracy of the segmentation result in the original image.

2. Materials and methods

2.1. Mammogram data sets

Two mammogram datasets, the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [29–31] and INBreast dataset [32] (**Table 1**), were used to train and test our mass detection and segmentation models. Each of the two datasets is composed of multiple pairs of scanned film mammography images (**Fig. 1(a)**) and their corresponding binary segmentation images of breast masses (**Fig. 1(b)**). The CBIS-DDSM has its standardized training-testing split as shown in **Table 1**. We used the training part of the CBIS-DDSM dataset as the training set in both of our mass detection and mass segmentation, and half of the CBIS-DDSM's test data is used for validation. In order to test the reliability and versatility of our model, we calculated its performance on the INBreast dataset and the other half of CBIS-DDSM's test data. It should be noted that each image in the two sources may have one or more than one breast mass.

2.2. Data pre-processing

DICOM (Digital Imaging and Communications in Medicine) is a special medical image file format, containing much detailed clinical information. In order to allow the object detection model and segmentation model to read the images directly, we converted the DICOM format to PNG format. We also found additional signals in the original image, which are useful for human reading, but may interfere with computer processing, such as text marks in the upper right corner of the original mammography images (**Fig. 1(a)**). To remove such non-informative noise and keep the size and/or orientation style consistent, we pre-processed all of our images using adaptive denoising algorithm (**Fig. 1**). We first did adaptive cropping, where we cropped 2.5% from the top and bottom of the original image. Then we binarized the images with a specific threshold value, then generated an adaptive mask by selecting the largest connected area in the binarized image using Python package scikit-image [33]. Only pixels within this adaptive mask were retained for further processing. Some images have the adaptive mask on the opposite side (i.e., the largest connected area could present on either the left side or right side of the original image). To make this consistent, we flipped some of the images to make the adaptive mask always on the left side, and once an image is flipped, this information will be recorded, its corresponding segmentation mask will also be flipped. After flipping, the contrast limited adaptive histogram equalization (CLAHE) were used to improve image contrast. CLAHE is a mature image processing algorithm and has been shown to have the ability to benefit the mammogram detection task [34]. Finally, we scaled the processed image to the default size by adaptive padding. We made sure that the cropping, flipping, and padding were operated based on the image size. Since the sizes of image and mask are the same in the input stage, we ensured that the adaptive operation did not damage the original labeled mass segmentation. Thus, our pre-processing algorithm guarantees the consistency of the masks and mammography images before and after the imaging processing. The whole processing is automatic, with two specify hyperparameters to be set: the threshold of binary the image and the ratio of imaging cropping. We did simple statistics and found that the gap between the breast body boundary and the image boundary is less than 2.5% for more than 90% of our mammograms. Therefore, the ratio of imaging cropping was set to 2.5%. The pre-processed images are the input of the YOLO-LOGO mass detection and segmentation model we proposed.

Table 1
Summary of the two datasets used in the study.

Dataset	Number of images with mask	Mammogram image	Segmentation mask
CBIS-DDSM [30]	1231(Training)+361(Test)	DICOM	DICOM
INBreast [32]	107	DICOM	PNG

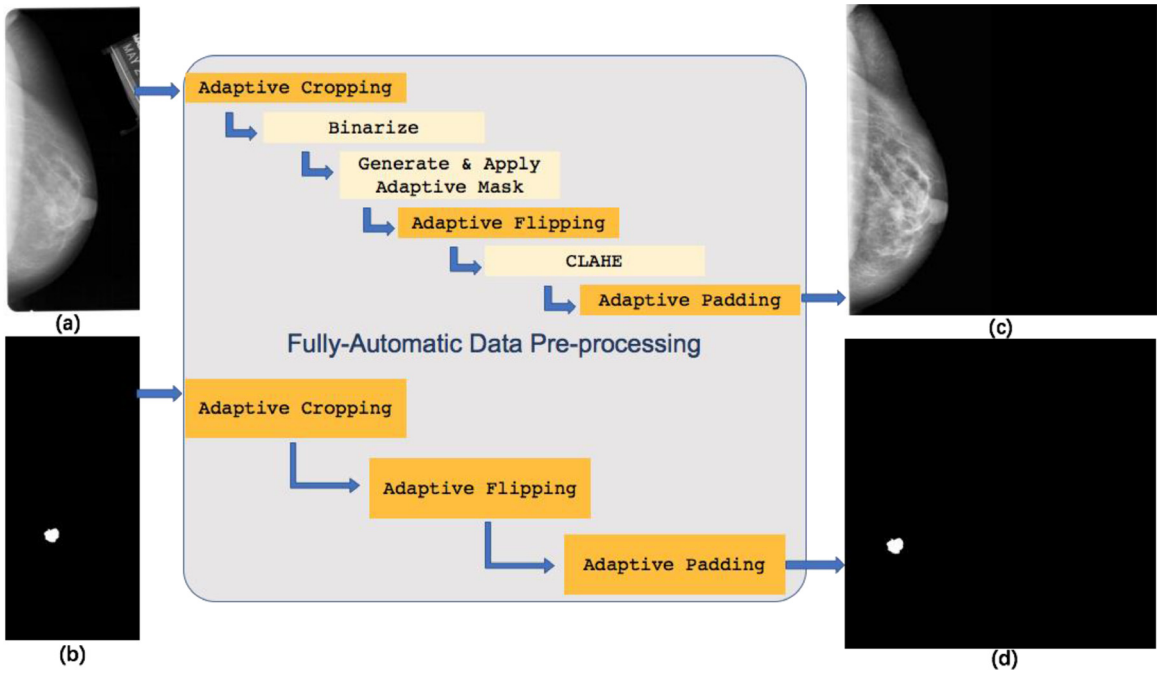


Fig. 1. The automatic adaptive denoising image pre-processing workflow. (a) is the original DICOM image with extraneous information on its upper right corner, (b) is the original ground truth segmentation mask of the same size corresponding to (a). (c) and (d) are the jpeg format mammography image and the ground truth segmentation mask after pre-processing, which have the same size (the default size is 4096×4096).

2.3. YOLO-LOGO model for breast mass detection and segmentation

The overall architecture of the proposed YOLO-LOGO transformer model for breast mass detection and segmentation in digital mammograms is shown in **Fig. 2(a)**. It includes two steps: Firstly, we use YoloV5 to detect the breast mass ROI, and then crop it directly from the high-resolution image (**Fig. 2(b)**); Secondly, to increase the training efficiency, we adopt an updated version of local-global (LOGO) segmentation strategy, which can greatly improve the segmentation resolution at the original pixel level (**Fig. 2(a)**).

2.3.1. YOLO: architecture and methods of breast mass detection

Unlike traditional multi-object detection tasks, our object is breast mass only (i.e., single class object detection). Thus, we do not need to classify the detected objects. In our breast mass detection stage, we adopted YoloV5L6, the state-of-the-art object detection model in computer vision, as the detector (**Fig. 2(b)**). YOLO is a classic object detection model and YoloV5L6 is its 5th version. YoloV5L6 contains three main substructures in its architecture: backbone, neck, and prediction. Backbone is used to extract features from the input data. YoloV5L6 uses the Cross Stage Partial Networks (CSP) [35] as the backbone. The extracted image features will then be passed to the model neck, which is used to generate features pyramids so that the model can detect the same object with different sizes and scales. YoloV5L6 uses Path Aggregation Network (PANet) [36] as the model neck. Then the feature pyramid created by PANet will be passed into the model head to generate the final output.

We used the enlarged bounding rectangles automatically extracted from the mask image as the ground truth ROI for YoloV5L6 object detection. Mathematically, these rectangles labels are stored as numeric values in the document. If there are n breast masses in the ground truth, there will be n lines in the label file, and each line represents one mass's positional and size information: relative coordinate X and Y of the center point, relative width (W) and relative height (H). Their mathematical definitions are as below.

$$relative\ coordinate\ X = \frac{x}{image\ width} \tag{1}$$

$$relative\ coordinate\ Y = \frac{y}{image\ width} \tag{2}$$

$$relative\ width\ W = \frac{mass\ width}{image\ width} \tag{3}$$

$$relative\ height\ H = \frac{mass\ height}{image\ height} \tag{4}$$

where x and y are the coordinates of the mass center, respectively. Mass width and mass heigh are measured from the breast mass. After the ROI labels were well-prepared, we further did image augmentations, including image rotation, scaling, horizontal flipping, vertical flipping, cropping, and mixing up, to enlarge the input data for each epoch of YoloV5L6 training. After 1000 epochs using the CBIS-DDSM training data, the loss of YoloV5L6 ROI detection model became stable and converged to a small value. When applying the trained mass detection model to identify the breast mass in an unlabeled mammogram, its output is the position and size of the mass in the original image with a confidence score. The higher the

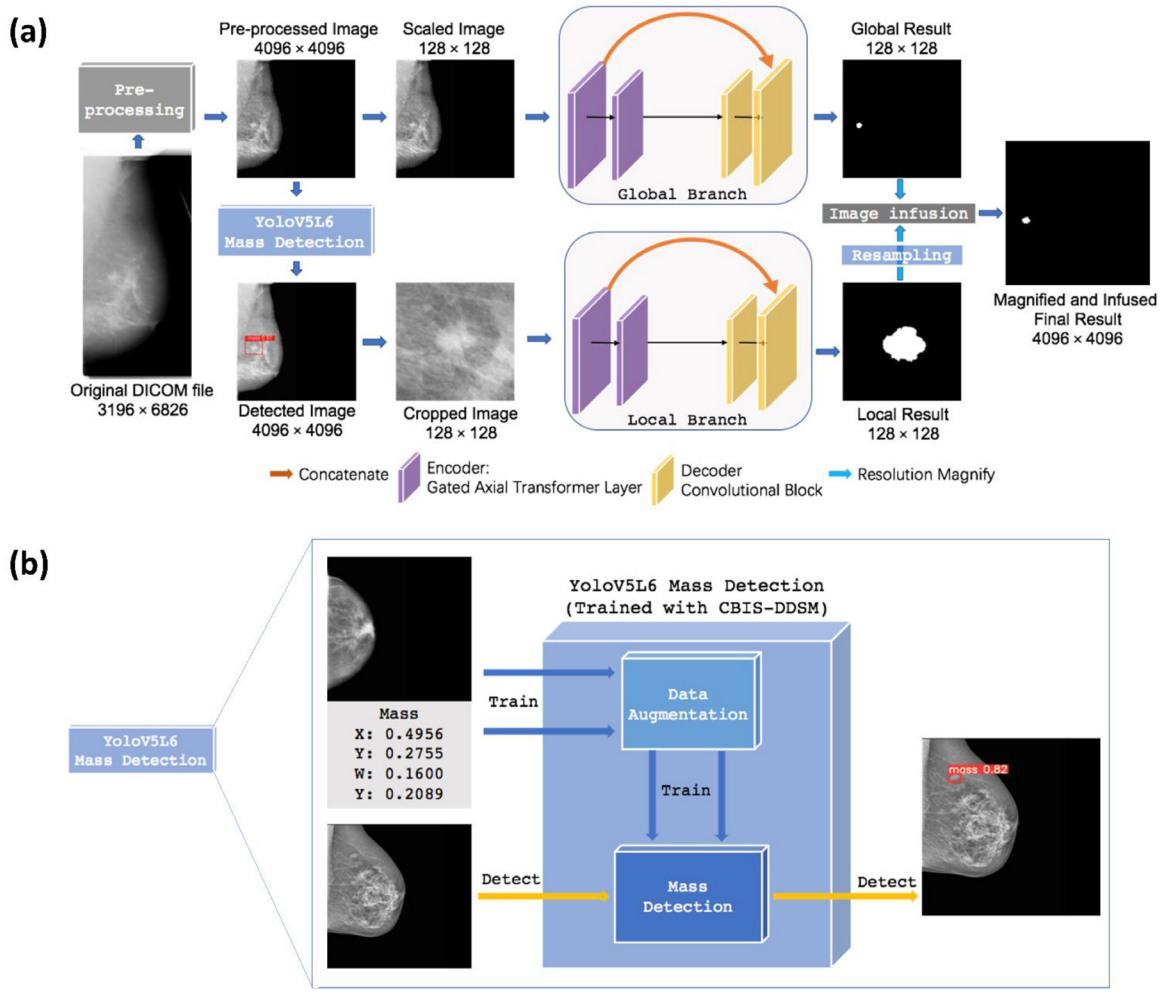


Fig. 2. The architecture of the proposed YOLO-LOGO model. **(a):** The overall workflow of this study with detailed LOGO structure of the proposed segmentation model. **(b):** The details of YOLO-based mass detection.

confidence score, the more likely it is a mass. More details of the YoloV5L6 could be found in its original paper [25,26].

2.3.2. LOGO: architecture and methods of breast mass segmentation

The next stage after mass detection is breast mass segmentation. We borrowed the gated axial-attention mechanism and LOGO training strategy from MedT [16] to achieve this. As mentioned briefly in the introduction, the gated axial-attention mechanism was developed to be better accepted when the sample size is not large enough (which is often the case in medical imaging field) and the computing resource is limited. The gated axial-attention could be considered as a variation or extension of self-attention, with gates added to control the information flow and the attention itself decomposed into two axials (height and width) to save computational costs. The self-attention mechanism proposed in the original ViT could be formulated as below:

$$o_{ij} = \sum_{h=1}^{H}\sum_{w=1}^{W} softmax\left(x_{ij}^{T} M_q^{T} M_k x_{hw}\right) M_v x_{hw} \tag{5}$$

Where $o \in \mathbb{R}^{D_{out} \times H \times W}$ is the output of self-attention layer. H, W, and D correspond to height, width, and dimension of input feature map $x \in \mathbb{R}^{D_{in} \times H \times W}$. x and o are high-dimensional matrices, $M \in \mathbb{R}^{D_{in} \times D_{out}}$ represents learnable projection weight matrix. $i, j \in N$ and $i \in [1, H]$, $j \in [1, W]$ represent coordinates of the input feature map x. While q, k and v are query, key, and value [35]. As can be seen from the Eq. 5, the self-attention layer could capture non-local information from the entire input feature map. Thus, it is computationally cost to train. To reduce the computational complexity, axial-attention was proposed to decompose the self-attention into two parts [37]. The first part calculates self-attention on height axis of the feature map (Eq. 6), while the second part performs on the width axis (Eq. 7).

$$o_{ij}^{H} = \sum_{h=1}^{H} softmax\left(x_{ij}^{T} M_q^{T} M_k x_{hj} + x_{ij}^{T} M_q^{T} r_{hj}^{q} + M_k^{T} x_{hj} r_{hj}^{k}\right)\left(M_v x_{hj} + r_{hj}^{v}\right) \tag{6}$$

$$o_{ij}^{W} = \sum_{w=1}^{W} softmax\left(x_{ij}^{T} M_q^{T} M_k x_{iw} + x_{ij}^{T} M_q^{T} r_{iw}^{q} + M_k^{T} x_{iw} r_{iw}^{k}\right)\left(M_v x_{iw} + r_{iw}^{v}\right) \tag{7}$$

where r^q, r^k, and r^v are learnable weight matrices for relative positional encodings for query, key, and value. These relative positional encodings usually need big data to train. However, medical image datasets usually do not have big sample size, which will harm the model performance because the learned relative positional encodings are not accurate. Therefore, in the medical image data analysis situation, it is better not to always add the learned relative positional encodings to the final output. Following this idea, gate

Table 2
Hyperparameter tuning.

	Parameter name	Parameter value
Breast mass detection (YoloV5L6)	Epoch	1E+2, 3E+2, 5E+2, 1E+3*
	Learning rate	1E-1, 1E-2, 1E-3
	Batch size	4, 8, 12, 16
	Patience	25, 50, 75
	Threshold	0.25, 0.3, 0.4, 0.45
	Optimizer	SGD, Adam
	Envolve	True
Breast mass segmentation (YOLO-LOGO)	Epoch	1E+2, 3E+2, 5E+2, 1E+3
	Learning rate	1E-4, 3E-4, 1E-3, 3E-3
	Batch size	1, 2, 4
	Image Size	128
	Augmentation	True
	Momentum	0.8, 0.9, 0.99
	Weight Decay	3E-6, 1E-5, 3E-5
	Optimizer	Adam, SGD

* Bold ones are used in the final model.

mechanism can be added to the query, key, and value of both height axial-attention (Eq. (8)) and width axial-attention (Eq. (9)) to control the information flow.

$$o_{ij}^{H} = \sum_{h=1}^{H} softmax\left(x_{ij}^{T} M_q^T M_k x_{hj} + G_q x_{ij}^T M_q^T r_{hj}^q + G_k M_k^T x_{hj} r_{hj}^k\right) \times \left(G_{v1} M_v x_{hj} + G_{v2} r_{hj}^v\right) \tag{8}$$

$$o_{ij}^{W} = \sum_{w=1}^{W} softmax\left(x_{ij}^{T} M_q^T M_k x_{iw} + G_q x_{ij}^T M_q^T r_{iw}^q + G_k x_{iw}^T M_k^T r_{iw}^k\right) \times \left(G_{v1} M_v x_{iw} + G_{v2} r_{iw}^v\right) \tag{9}$$

G_q, G_k, G_{v1}, G_{v2} are learnable gate parameters, which are used to control the pass of learned relative position codes to the final output.

The ROI detected by the YoloV5L6 model in the previous step is a local view of the breast mass that could provide zoomed-in details of the breast mass, while the whole image could provide long-range non-local context. Both ROI and whole image are important for archiving a good segmentation performance. We used the LOGO architecture to take advantage of both ROI and whole image for our final segmentation result. The LOGO architecture has two branches, local branch and global branch (**Fig. 2B**). After breast mass detection is completed, we can get masses' relative coordinates and size from the output of YoloV5L6. Then we crop the squares of the masses from the high-resolution raw images. The cropped images preserve as much detailed information as possible for local context, thereby generating more refined segmentation results in this local branch. At the same time, we used the pre-processed breast mammogram images with reduced resolution (128×128) as the input of the global segmentation branch. This is to maintain the position accuracy of the final segmentation result. Both the global branch and the local branch are composed of the same number of gated axial transformer layers. After concurrent local and global segmentation, we strictly abide by the coordinate and size information and infuse the generated local segmentation into the generated global segmentation. This generates the final mass segmentation result for a full mammography image.

2.3.3. Hyperparameter finetuning

The hyperparameters we tuned for training our breast mass detection model and breast mass segmentation model are listed in **Table 2**. Please noticed that the bold ones were used in the final model as they could provide the lowest loss values and the most stable training procedure.

2.4. Baselines and performance metrics

For the breast mass detection, we compared the performance of our proposed YOLO model with several object detection baselines such as Faster Region-based CNN (R-CNN) [38–40], Single Shot Detector (SSD) [41,13] and other YOLO versions [23,42–44]. These are representative object detection models. Faster R-CNN solves the object detection problem by using a method called selective search which reduces the computational burden caused by sliding window in traditional object detection models. It has been shown to have a decent performance in mammogram-based breast mass detection [45]. SSD runs the convolutions at different scales and each scale could output detection bounding box in different sizes. The performance of these models was also discussed and compared in previous works [23]. We evaluated the performance using true positive rate (TPR) [46] and mean average precision (mAP) [47].

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$mAP = \frac{1}{n} \sum_{n} \frac{\sum_{r}^{R} Precision@r}{R} \tag{11}$$

TP, *FP*, and *FN* are the true positives, false positives, and false negatives, respectively. AP is defined as the averaged precision among several selected recalls. *Precision@r* is the precision when recall is *r*; *R* is the number of selected recalls. *n* is the number of classes in object detection (particularly in breast mass detection, $n = 1$). Recall and precision are defined as:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

For breast mass segmentation, we compared the performance of our proposed LOGO model with several baseline methods such as Gated Axial Net [37] and MedT [16]. As mentioned in the introduction, MedT proposed a gated axial transformer model structure which consists of convolutional layers and gated axial-attention layers with local and global branches but without YOLO detection [16]. Gated Axial Net consists of only gated axial-attention layers [37]. To evaluate the effect of the local-global design on the model's performance, we applied the Gated Axial Net in both local way and global way: Gated Axial Net (Global) is based on the whole image only. The Gated Axial Net (Local) is based on the YOLO detected ROIs only. We evaluated and compared their performance using F1-score and IoU (also known as Jaccard index),

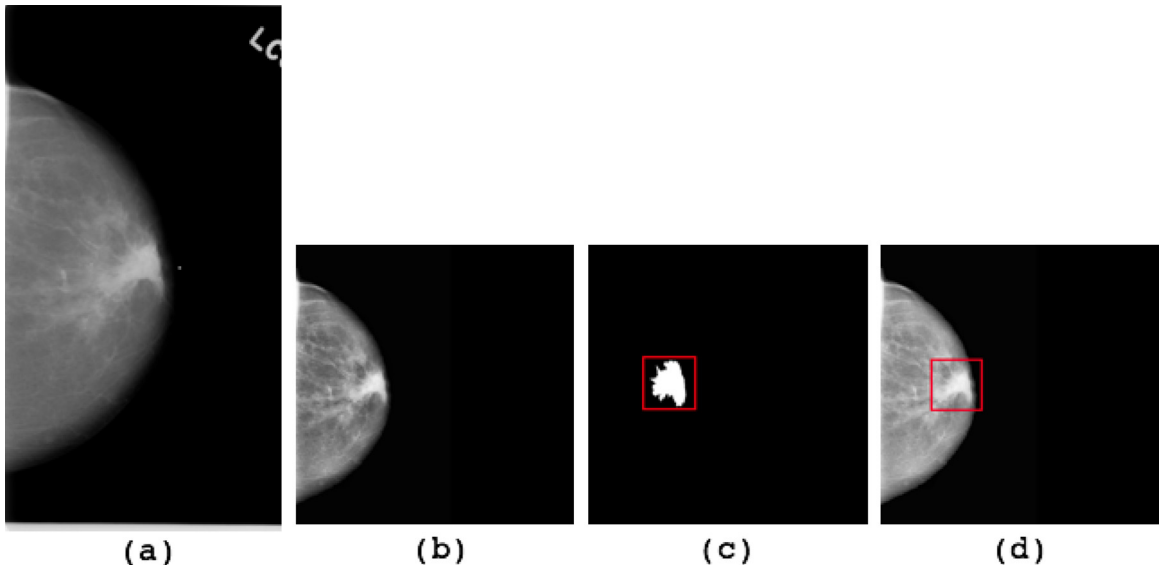


Fig. 3. Automatically extracting the ground truth label for mass detection. (a): An example of original image. (b): The preprocessed image of (a) using adaptive denoising algorithm. (c): a bounding rectangle(s) of mass region(s) is created on the pre-processed binary mask image of (b). (d): The same-sized rectangle(s) is positioned on the same position of (b), this forms the ground truth label for training YoloV5L6.

which are defined as below.

$$F1 = \frac{2Precision * Recall}{Precision + Recall} \tag{14}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{15}$$

2.5. Sensitivity analysis of the proposed YOLO-LOGO model

We investigated the potential effects of image resolutions and augmentations on the performance of the proposed YOLO-LOGO model on the CBIS-DDSM test set and INBreast dataset. For the data augmentation, we increased the data size to the original two times using a parametric model developed by ThambawitaVajira et al. [48]. For the resolutions, we considered the resolution sizes of 64×64, 128×128 and 256×256 with and without data augmentation.

3. Results

3.1. Data pre-processing

As we have shown in **Fig. 3(a)**, the original mammography images often contain some irrelevant information which has been removed in **Fig. 3(b)**. It is also noted that the file format is changed from DICOM to JPEG, which reduced the file size by 90% while preserving the details of the image. After pre-processing, we got enhanced JPEG images of uniform size 4096×4096. It is also noted that we've adapted the same automatic adaptive procedures for the two datasets (CBIS-DDSM and INBreast) used in our study.

3.2. YOLO-based breast mass detection

We trained and tested several variants of YoloV3 and YoloV5, and their breast mass detection performances on the two datasets are listed in **Table 3**. In YoloV5′s 6th updates, YoloV5s6, YoloV5m6, and YoloV5L6 are able to process input images with a higher resolution of 1280 × 1280. Although containing the most trainable parameters (77 M), YoloV5L6 achieved the best mAP performance on both of the datasets. Thus, it was selected as the breast mass detection model in our study.

Table 3
The test performance of variants of YoloV3 and YoloV5. YoloV5L6, which could be trained with high resolution (1280×1280) data, achieved the best mAP performance on both datasets. It should be noted that the models were trained on CBIS-DDSM's training set only. The reported performance is measured on CBIS-DDSM's test set. For INBreast data, we only consider it as another test set. No training was conducted on INBrest data.

Models	mAP on test set		Parameters	Image size
	CBIS-DDSM	INBreast		
YoloV3-tiny:	53	45.1	8.9M	640
YoloV3	60	56.5	61.9M	640
YoloV3-spp	63	58.3	63.0M	640
YoloV5s	59	47.2	7.3M	640
YoloV5L	60	53.0	47.0M	640
YoloV5s6	58	53.2	12.7M	1280
YoloV5m6	60	57.7	35.9M	1280
YoloV5L6	65	61.4	77.2M	1280

Five examples of the YoloV5L6's output with the detected bounding boxes and confidence scores are shown in the first row of **Fig. 4**. Their ground truth labels are showed in the corresponding columns of the second row of **Fig. 4**. **Fig. 4(a)–(c)** are 3 successful cases with confidence scores equal to 0.9. Comparing with their ground truth in **Fig. 4(f)–(h)**, the bounding boxes are positioned on the correct place and the size of the bounding boxes are roughly consistent with the ground truth labels. **Fig. 4(d)** shows a detected ROI bounding box with a confidence score only equals to 0.4. Although the detected bounding box appears to be correct compared to its ground truth (**Fig. 4(i)**), the model has no confidence in the prediction. This might be due to the high density of breast tissue and the small area of the breast itself in this example. **Fig. 4(e)** shows a failure case where the model failed to output a bounding box because the confidence score is below the specified threshold of 0.4. This may be because the breast mass is too small and hidden in normal breast tissues. Images like this without object detection output are directly used as input to the global branch in the following segmentation stage. Correspondingly, the result of such segmentation does not require infusion (as there will be no local branch in this case).

SSD and Faster R-CNN were acted as baselines of the breast mass detection model in this study. **Fig. 5(a)** visualized their breast

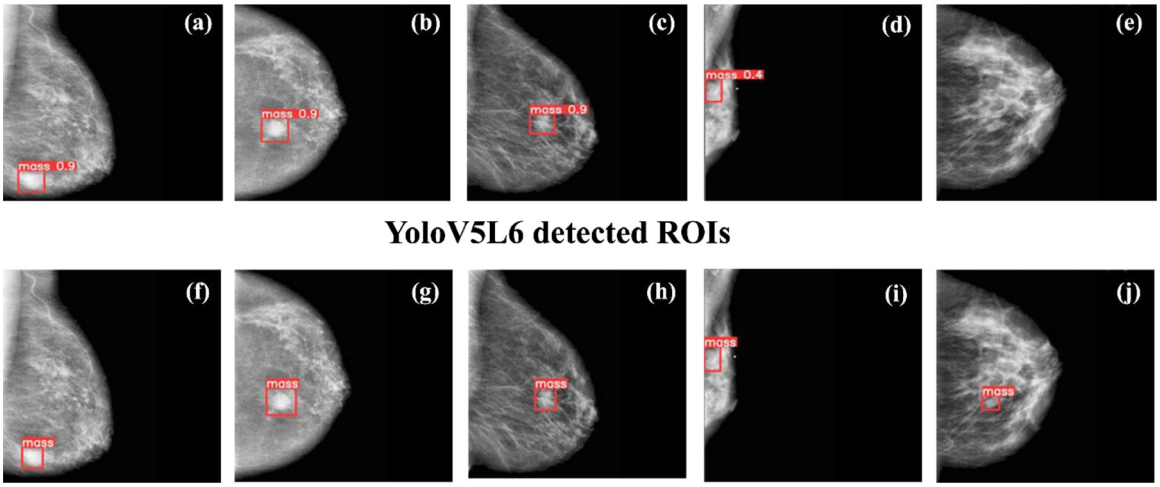
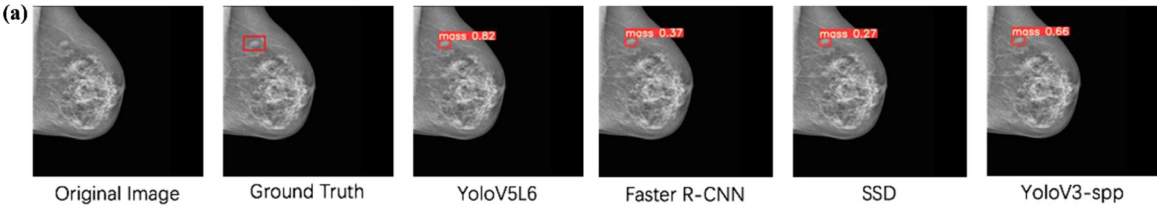


Fig. 4. The success and failure cases of breast mass detection using YoloV5L6 model. **(a, b, c)** are 3 success cases with confidence scores equal to 0.9. Their ground truth labels are shown in **(f, g, h),** respectively. **(d)** is a relative failure case with confidence score equal to 0.4. Its ground truth is shown in **(i). (e)** is a failure case, where the model fails to output a ROI bounding box. Its ground truth is **(j).**



(b)

Models	mAP on Dataset		TPR on Dataset	
	CBIS-DDSM	INBreast	CBIS-DDSM	INBreast
YoloV5L6 (Pytorch)	**65**	**61.4**	**95.7**	**92.5**
Faster R-CNN (Tensorflow)	50.6	47.0	84.0	79.8
Single-Shot Detection (Pytorch)	48.3	42.1	86.1	82.7
YoloV3-spp (Pytorch)	63	58.3	93.4	88.2

Fig. 5. The performance comparisons of the baseline breast mass detection models and the proposed YoloV5L6. **(a)** visualized the breast mass detection results of the models on an example image. **(b)** is the performance metrices of the baseline models compared with the best variant of YoloV3 (YoloV3-spp) and YoloV5 (YoloV5L6). It should be noted that the proposed model and the baseline models were trained and tested on the same training set and validation set from CBIS-DDSM data only. The reported performance is measured on CBIS-DDSM's test set. For INBreast data, we only consider it as another test set. No training was conducted on INBrest data.

mass detection result on an example image. Although there were some discrepancies between the detected ROI and ground truth ROI. Regarding the size of the bounding box, YoloV5L6 achieved the highest confidence score and the accurate ROI position. The performance metrices of the baseline models were compared with the best variant of YoloV3 and YoloV5 in **Fig. 5(b)**. The training was done on Nvidia GeForce GTX 1080Ti GPU. During the training, the epoch was set to be 1000 and batch size was set to be 4. YoloV3-spp and YoloV5L6 held significantly higher scores and more stable performances on both of the datasets than the SSD and Faster R-CNN models while YoloV5L6 was slightly better than YoloV3-spp.

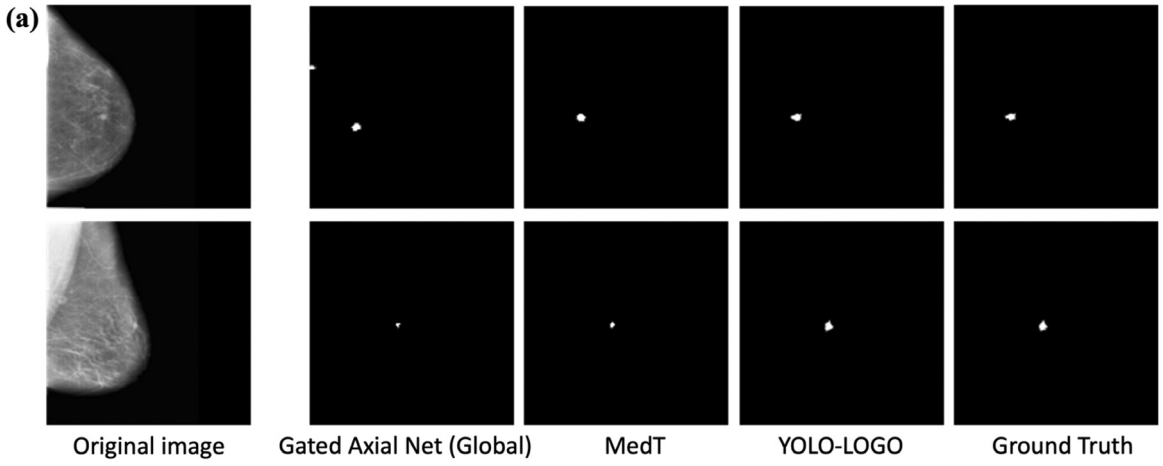
3.3. YOLO-LOGO based breast mass segmentation

Gated Axial Net and MedT were used as baselines of the breast segmentation model YOLO-LOGO in this study. Two example segmentations were visualized in **Fig. 6(a).** The proposed YOLO-LOGO best preserved the shape and position information of the breast masses compared with the Gated Axial Net and MedT. In the first case (the first row), Gated Axial Net incorrectly predicted the breast mass position (lower than the ground truth, and an extra piece at the left upper corner was predicted as mass although it was not). In the second case (the second row), although Gated Axial Net and MedT correctly predicted the location of the mass, their segmentations are not as good as YOLO-LOGO because they tend to underestimate the area of the mass. The F1 and IoU metrices are shown in **Fig. 6(b).** YOLO-LOGO outperformed the previous works on both datasets and the local resolution is also preserved.

4. Discussion

In this study, we used an automatic adaptive denoising image pre-processing framework for mammography data. We proposed the state-of-the-art object detection model, YoloV5L6, for the breast mass detection. Overall, the proposed YOLO-LOGO model outperformed other baselines in segmenting the breast masses using the mammography.

We investigated the potential effects of image resolutions and augmentations on the performance of the proposed YOLO-LOGO model on the CBIS-DDSM test set and INBreast dataset. As shown in **Table 4**, when the image resolution is 128×128 without augmentation, the model achieved the best F1 score (74.52), the third best IoU score (64.04) on CBIS-DDSM test set, and the second best F1 score (69.37), the second best IoU score (61.09) on INBreast



(b)

Model	CBIS-DDSM		INBreast		Image Size
	F1	IoU	F1	IoU	
Gated Axial Net (Global)	69.30	60.63	64.25	59.48	128*128
Gated Axial Net (Local)	72.06	61.79	60.73	54.61	128*128
MedT	70.18	60.30	65.50	53.58	128*128 (Global) 32*32 (Local)
YOLO-LOGO	**74.52**	**64.04**	**69.37**	**61.09**	128*128 (Both global and local)

Fig. 6. The performance comparisons of the baseline breast mass segmentation models and the proposed YOLO-LOGO. (a) The segmentation results on two test examples. In the first row, the result of Gated Axial Net incorrectly predicted the mass position. In the second row, although Gated Axial Net and MedT correctly predicted the location of the mass, they tend to underestimate the area of the mass. As comparison, YOLO-LOGO best predicted the size, position, and shape of the mass. **(b)** The comparisons of the segmentation performance metrices on two datasets. Gated Axial Net (Global) is based on the whole image only. The Gated Axial Net (Local) is based on the YOLO detected ROIs only. The MedT is based on global and local without YOLO detection. YOLO-LOGO outperformed the previous works on both datasets and the local resolution is also preserved.

Table 4
The sensitivity analyses of the proposed YOLO-LOGO model.

Image resolution	CBIS-DDSM		INBreast	
	F1	IOU	F1	IOU
64×64	72.32	**65.90**	68.23	**61.25**
64×64 (with augmentation)	69.85	64.10	62.69	59.09
128×128	**74.52**	64.04	69.37	61.09
128×128(with augmentation)	71.96	63.92	65.98	60.40
256×256	73.94	62.87	**69.41**	60.52
256×256(with augmentation)	72.21	63.08	64.74	57.78

dataset. Overall, the best performance has been achieved using the data set with resolution 128 * 128 without augmentation. Although it is expected that augmentation will likely increase the performance of the model under the assumption that the training data and the testing data are both drawn from the same distribution, our study sample size is relatively small, which may not represent the distribution of the training set. Therefore, the augmentation technique applied here does not show expected performance. Thus, to balance the computational cost and the model performance, we decided to us the image resolution of the proposed YOLO-LOGO model as 128×128 without augmentations.

Unlike the two-stage model proposed by Yan et al., which could only output one breast mass per image, we set a threshold for our YOLO-LOGO segmentation model. Those detected ROIs with confidence scores higher than this threshold would be recognized as a mass in output. Thus, it is not limited to only one mass per image. Consequently, the detected ROI could be 0 for some images. In this case, the whole image will be input into the global branch of our YOLO-LOGO model while the local branch left empty, which means the final segmentation result will only depend on the whole image processed by the global branch. This makes our model more flexible and suitable for mammography data as the breast masses are easy to be missed by the detection model due to the small mass size and high tissue density.

One limitation of our work is that the proposed model is not trained end-to-end. The YoloV5L6 based breast mass detection model and the gated axial transformer plus LOGO training strategy-based segmentation model are trained separately. Thus, in the future, we could consider combining their loss functions into one and making the system end-to-end. Also, it should be noted that although our proposed model and the baseline models used in this study were trained, validated, tested on the same training set, validation set, and test set, it needs to be caution when comparing our model's performance with other published works that used the same CBIS-DDSM data but were not included in our baselines. Because we did not use the CBIS-DDSM provided test set. Instead, we divided the original test set into test and validation sets. The validation set is critical for the hyperparameters' finetuning in our model. Since the INbreast data are full-field digital mammography while the CBIS-DDSM are scanned films, they may have been acquired using different imaging modalities. Our models were trained on CBIS-DDSM data only, and the INBreast (as an independent dataset) acted as an additional test set to evaluate the generality of our model. We observed that our model's performance on the CBIS-DDSM test set is better than that on the INBreast, which maybe partially due to the acquisition modality difference.

5. Conclusions

We developed a two-stage model to first detect breast mass, and then solve the breast mass segmentation problem. The data was carefully pre-processed, and the irrelevant noise information was removed from the images, bringing additional accuracy for

later analysis. Our model showed the improved performance than state-of-the-art models in both breast mass detection and segmentation. Additionally, this model does not require high powerful computing resources. It could run on ordinary accessible devices. Thus, it is of practical significance to promote the model to the precise CAD for mammography.

Funding

This work was supported in part by CancerCare Manitoba Foundation, Research Manitoba and Mitacs. P.H. is the holder of Manitoba Medical Services Foundation (MMSF) Allen Rouse Basic Science Career Development Research Award.

Declaration of Competing Interest

All authors declared no conflict of interest.

CRediT authorship contribution statement

Yongye Su: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Qian Liu:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Wentao Xie:** Conceptualization, Writing – review & editing. **Pingzhao Hu:** Conceptualization, Methodology, Funding acquisition, Resources, Writing – review & editing, Supervision.

Acknowledgments

We thank for Dr. Inês Domingues to share the INBreast Dataset used in the study. We also thank for Dr. Daniel L. Rubin and The Cancer Imaging Archive to make the CBIS-DDSM data publicly available.

References

[1] P. Boyle, B. Levinothers, World Cancer Report 2008, IARC Press, International Agency for Research on Cancer, 2008.
[2] D.M. Parkin, L.M.G. Fernández, Use of statistics to assess the global burden of breast cancer, Breast J. 12 (2006) S70–S80.
[3] P.C. Gøtzsche, M. Nielsen, Screening for breast cancer with mammography, Cochrane Database of Syst. Rev. (2009), doi:10.1002/14651858.CD001877.pub3.
[4] M.G. Marmot, D.G. Altman, D.A. Cameron, J.A. Dewar, S.G. Thompson, M. Wilcox, The benefits and harms of breast cancer screening: an independent review, Br. J. Cancer 108 (2013) 2205–2240, doi:10.1038/bjc.2013.177.
[5] J. Wei, H.P. Chan, C. Zhou, Y.T. Wu, B. Sahiner, L.M. Hadjiiski, M.A. Roubidoux, M.A. Helvie, Computer-aided detection of breast masses: four-view strategy for screening mammography, Med. Phys. 38 (2011) 1867–1876, doi:10.1118/1.3560462.
[6] O. Ronneberger, P. Fischer, T. Brox, N. Nassir, J. Hornegger, U-Net: convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 W.W.M. and F.A.F., Springer International Publishing, Cham, 2015, pp. 234–241.
[7] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: a Nested U-Net architecture for medical image segmentation, Deep Learning in Medical Image Analysis and Multimodal Learning For Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction With MICCAI, Granada, Spain, 2018, pp. 3–11. S... 11045, doi:10.1007/978-3-030-00889-5_1.
[8] Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: learning where to look for the pancreas.
[9] X. Li, H. Chen, X. Qi, Q. Dou, C.W. Fu, P.A. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (2018) 2663–2674, doi:10.1109/TMI.2018.2845918.
[10] M.Z. Alom, C. Yakopcic, T.M. Taha, V.K. Asari, Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net), in: Proceedings of the NAECON 2018 - IEEE National Aerospace and Electronics Conference, 2018, pp. 228–233, doi:10.1109/NAECON.2018.8556686.
[11] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.W. Chen, J. Wu, UNet 3+: a full-scale connected UNet for medical image segmentation, in: Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 1055–1059, doi:10.1109/ICASSP40776.2020.9053405.
[12] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, A.S. Elmaghraby, Connected-UNets: a deep learning architecture for breast mass segmentation, NPJ Breast Cancer 7 (1) (2021) 1–12, doi:10.1038/s41523-021-00358-x.
[13] V.K. Singh, H.A. Rashwan, S. Romani, F. Akram, N. Pandey, M.M.K. Sarker, A. Saleh, M. Arenas, M. Arquez, D. Puig, J. Torrents-Barrena, Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network, Expert Syst. Appl. 139 (2020), doi:10.1016/j.eswa.2019.112855.
[14] K.B. Soulami, N. Kaabouch, M.N. Saidi, A. Tamtaoui, Breast cancer: one-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation, Biomed. Signal Process. Control 66 (2021), doi:10.1016/j.bspc.2021.102481.
[15] L. Tsochatzidis, P. Koutla, L. Costaridou, I. Pratikakis, Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses, Comput. Methods Programs Biomed. (2021) 200, doi:10.1016/j.cmpb.2020.105913.
[16] Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M., 2021. Medical transformer: gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662.
[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: transformers for image recognition at scale.
[18] Chefer, H., Gur, S., Wolf, L., 2020. Transformer interpretability beyond attention visualization.
[19] J. Clauwaert, G. Menschaert, W. Waegeman, Explainability in transformer models for functional genomics, Brief. Bioinform. 22 (2021), doi:10.1093/bib/bbab060.
[20] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-Unet: unet-like pure transformer for medical image segmentation.
[21] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L., 2021. CvT: introducing convolutions to vision transformers.
[22] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: transformers make strong encoders for medical image segmentation.
[23] Y. Yan, P.H. Conze, G. Quellec, M. Lamard, B. Cochener, G. Coatrieux, Two-stage multi-scale breast mass segmentation for full mammogram analysis without user intervention, Biocybern. Biomed. Eng. 41 (2021) 746–757, doi:10.1016/j.bbe.2021.03.005.
[24] Redmon, J., Farhadi, A., 2018. YOLOv3: an incremental improvement.
[25] Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, Changyu, L., V, A., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Hajek, J., Diaconu, L., Marc, Kwon, Y., oleg, wanghaoyang0106, Defretin, Y., Lohia, A., ml5ah, Milanko, B., Fineran, B., Khromov, D., Yiwei, D., Doug, Durgesh, Ingham, F., 2021a. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, supervise.ly and YouTube integrations. doi:10.5281/ZENODO.4679653.
[26] Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG, Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M.T., Marc, albinxavi, fatih, oleg, wanghaoyang0106, 2021b. ultralytics/yolov5: v6.0 - YOLOv5n "nano" models, Roboflow integration, TensorFlow export, OpenCV DNN support. doi:10.5281/ZENODO.5563715.
[27] Cheng, H., Lian, D., Deng, B., Gao, S., Tan, T., Geng, Y., 2019. Local to global learning: gradually adding classes for training deep neural networks.
[28] Jeub, L.G.S., Colavizza, G., Dong, X., Bazzi, M., Cucuringu, M., 2021. Local2Global: scaling global representation learning on graphs via local training.
[29] R.S. Lee, F. Gimenez, A. Hoogi, D.L. Rubin, Curated breast imaging subset of DDSM, Cancer Imaging Arch. (2016), doi:10.7937/K9/TCIA.2016.7O02S9CY.
[30] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data 4 (2017) 170177, doi:10.1038/sdata.2017.177.
[31] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digit. Imaging 26 (2013) 1045–1057.
[32] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, INbreast, Acad. Radiol. 19 (2012), doi:10.1016/j.acra.2011.09.014.
[33] S. Van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, scikit-image: image processing in python, Peer J. 2 (2014), doi:10.7717/peerj.453.
[34] E.D. Pisano, S. Zong, B.M. Hemminger, M. DeLuca, R.E. Johnston, K. Muller, M.P. Braeuning, S.M. Pizer, Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms, J. Digit. Imaging 11 (4) (1998) 193–200.
[35] C.Y. Wang, H.Y.M. Liao, Y.H. Wu, P.Y. Chen, J.W. Hsieh, I.H. Yeh, CSPNet: a new backbone that can enhance learning capability of CNN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391.
[36] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[37] Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T., 2019. Axial attention in multidimensional transformers.

[38] S. Famouri, L. Morra, L. Mangia, F. Lamberti, Breast mass detection with faster R-CNN: on the feasibility of learning from noisy annotations, IEEE Access 9 (2021) 66163–66175, doi:10.1109/ACCESS.2021.3072997.
[39] J. Hung, A. Carpenter, Applying faster R-CNN for object detection on malaria images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.
[40] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149, doi:10.1109/TPAMI.2016.2577031.
[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 21–37.
[42] M.A. Al-antari, S.M. Han, T.S. Kim, Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms, Comput. Methods Programs Biomed. 196 (2020) 105584, doi:10.1016/j.cmpb.2020.105584.
[43] M.A. Al-antari, M.A. Al-masni, M.T. Choi, S.M. Han, T.S. Kim, A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inform. 117 (2018) 44–54, doi:10.1016/j.ijmedinf.2018.06.003.
[44] M.A. Al-masni, M.A. Al-antari, J.M. Park, G. Gi, T.Y. Kim, P. Rivera, E. Valarezo, M.T. Choi, S.M. Han, T.S. Kim, Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system, Comput. Methods Programs Biomed. 157 (2018) 85–94, doi:10.1016/j.cmpb.2018.01.017.
[45] R. Agarwal, O. Díaz, M.H. Yap, X. Llado, R. Marti, Deep learning for mass detection in full field digital mammograms, Comput. Biol. Med. 121 (2020) 103774.
[46] F. Provost, R. Kohavi, Glossary of terms, J. Mach.Learn. 30 (1998) 271–274.
[47] MAP S.M. Beitzel, E.C. Jensen, O. Frieder, L. Liu, M.T. Özsu, Encyclopedia of Database Systems Springer US, Boston, MA, 2009, pp. 1691–1692, doi:10.1007/978-0-387-39940-9_492.
[48] ThambawitaVajira, V., Salehi, P., Sheshkal, S.A., Hicks, S.A., Hammer, H.L., Parasa, S., Lange, T., Halvorsen, P., Riegler, M.A., 2021. SinGAN-Seg: synthetic training data generation for medical image segmentation. arXiv:2107.00471