IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 March 2024, accepted 9 May 2024, date of publication 17 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402260

RESEARCH ARTICLE

Local Cross-View Transformers and Global Representation Collaborating for Mammogram Classification

WENNA WU, QI RONG, AND ZHENTAI LU, (Member, IEEE)
School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China
Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou 510515, China

Corresponding author: Zhentai Lu (13422322117@163.com)

This work was supported in part by the Science and Technology Planning Project of Guangdong Province under Grant 2020A1414040021, and in part by the Science and Technology Planning Project of Guangzhou City under Grant 202103000037.

ABSTRACT When analyzing screening mammography images, radiologists compare multiple views of the same breast to help improve the detection rate of lesions and reduce the incidence of false-positive results. Therefore, to make the deep learning-based mammography computer-aided detection/diagnosis (CAD) system meet the radiologists' requirements for accuracy and generality, the construction of deep learning models needs to mimic manual analysis and consider the correlation between different views of the same breast. In this paper, we propose the Local Cross-View Transformers and Global Representation Collaborating for Mammogram Classification (LCVT-GR) model. The model uses different view images to train in an end-to-end manner. In this model, the global and local representations of mammogram images are analyzed in parallel using the global-local parallel analysis method. To validate the effectiveness of our method, we conducted comparison experiments and ablation experiments on two publicly available datasets, Mini-DDSM and CMMD. The results of the comparison experiments show that our method achieves better results compared with existing advanced methods, with greater improvements in both AUC-ROC and AUC-PR assessment metrics. The results of the ablation experiments show that our model architecture is scientific and effective and achieves a good trade-off between computational cost and model performance.

INDEX TERMS Deep learning, mammogram classification, multi-view, cross-view transformers, global-local analysis.

I. INTRODUCTION

According to the WHO (World Health Organization), breast cancer has surpassed lung cancer as the most frequent cancer and the fifth largest cause of cancer mortality worldwide. Globally, there will be 2.3 million cases of breast cancer in women alone in 2020, and 685,000 people will pass away from the disease [1]. Breast cancer mortality has decreased by 40% in high-income countries since regular mammography screening was introduced by health authorities in the 1980s for age groups deemed to be at risk, in contrast to the situation in low- and middle-income countries [2]. Therefore, reducing breast cancer mortality globally requires early diagnosis and treatment of the disease. Since the pathogenesis of breast cancer is still unknown and there are currently no proven preventative measures, early diagnosis is still the best medical course of action [3].

Currently, the early diagnosis of breast cancer requires specialized radiologists and mammologists, which makes mammography screening programs costly to implement and can be more difficult to implement in countries with low incomes and a shortage of radiologists [2]. A few false positives can result from mammography screening, which can cause patients and their families unneeded worry and anxiety, additional imaging tests, and occasionally needle biopsies [4]. In contrast, deep learning-based AI-assisted

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

technology can streamline radiologists' evaluation of screening mammography images, increasing their effectiveness and precision. Because of this, deep learning has become a common technique for creating mammography computer-aided detection/diagnosis (CAD) schemes [5].

Two views of each breast are typically obtained during a mammogram: a top-down view known as craniocaudal (CC) and a lateral view known as mediolateral oblique (MLO). Radiologists search for certain abnormalities on mammograms to determine whether they are abnormal, the most frequent ones being masses, calcifications, structural deformities, and asymmetric densities [6]. The radiologist will refer to as many views as possible when looking for abnormalities to determine if a suspicious lesion is present. Figure 1 shows an example of a benign and malignant breast lesion, where the border features of a malignant breast lesion are distinctly different from those of a benign breast lesion. Most of the benign lesions have smooth, lobulated borders, whereas the malignant lesions have irregular, burr-like borders [7]. Comparing multiple views of the same breast can help improve the detection rate of lesions and reduce the incidence of false positive results [8], [9]. Therefore, we consider the correlation between different views of the same breast and construct a global-local analysis method with different views.

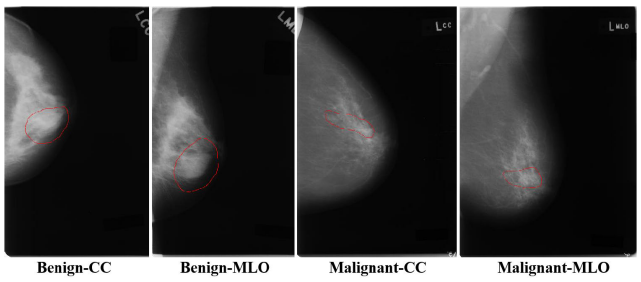

FIGURE 1. Examples of benign and malignant breast lesions.

The global-local analysis method combines features of the whole image with features of small localized blocks for classification. Global analysis helps to detect abnormalities such as distortion and asymmetric denseness of breast structures, and local analysis helps to detect abnormalities such as masses and calcifications. There are two mainstream global-local analysis methods, one is to extract global features from the whole image, and local features from within the region of interest, and then combine the extracted global and local features for classification [10], [11]. The other is to first train a patch classifier and later fine-tune it for the whole mammogram image classification [12]. Although the performance demonstrated by these two methods has been comparable to that of medical experts, there are still some shortcomings. The local features are too dependent on the accuracy of the region of interest localization, which may result in obtaining suboptimal solutions. In response, many scholars have proposed improved global-local analysis methods. Petrini et al. [13], [14] proposed an architecture consisting of multiple CNN (convolutional neural network) paths, where each path extracts features from different views, and then the output of the features from all paths are stitched together and finally sent to the fully connected layer for classification. Chen et al. [15] proposed a pure transformer model with local and global blocks to learn the dependencies between different views of the mammary gland. Although the improved model described above uses the complementary information of different views to learn the dependencies between different views, it does not explore the cross-view information. In our opinion, the global features of the same mammary gland are similar and the local features should differ. When extracting local features, making cross-attention between different view features is beneficial to find the interrelationship between lesions and extract more effective characterization information.

Cross-attention of different views can be achieved by cross transformers. The cross-attention mechanism of the cross transformer can guide the transformer to learn the association information of different features during the training process and achieve the effective fusion of different features [16]. Currently, cross transformers have been widely used in many fields. In the time-domain speech enhancement task, Wang et al. [16] used cross transformers to fuse local features extracted by local transformers and global features extracted by global transformers to obtain a better contextual feature representation, where the Q(Query) and K(Key), V(Value) input to the cross transformer are from different features. In a few-sample target detection task, Han et al. [17] achieved asymmetric batch cross-attention across branches by aggregating K and V from different features. In hyperspectral and multispectral image fusion tasks, Wang et al. [18] added the cross-attention idea to the traditional transformer self-attention mechanism to achieve information fusion between two modalities by exchanging K from different modal features. In the face recognition task, Li et al. [19] used cross transformers, which can remove the noisy information due to race while retaining useful identity features by exchanging V from different features.

In this paper, we propose the Local Cross-View Transformers and Global Representation Collaborating for Mammogram Classification (LCVT-GR) model. The model uses different view images to train in an end-to-end manner. In this model, the global representation and local representation of mammogram images are analyzed in parallel using the global-local parallel analysis method. In generating the local representation, a cross-view transformer is used to achieve information exchange between different view features. Finally, the classifier fuses the local representation and the global representation for the final prediction. The innovations of this paper are:

1) An improved global-local parallel analysis method for multi-view mammography images is proposed to analyze the global representation and local representation of the mammary gland in parallel.

2) A new local cross-view transformer is proposed to learn the dependencies between different views and achieve the information fusion between different views.

II. MATERIALS AND METHODS

A. DATA COLLECTION

Two open-source digital mammography image datasets, the MiniDigital Dataset for Screening Mammography (Mini-DDSM) [20] and the Chinese Mammography Database (CMMD) [21], were used in this study. The Mini-DDSM included 3904 breasts (1952 patients) from multiple centers, 1342 breast biopsies were confirmed benign, 1358 breast biopsies were confirmed malignant, and 1204 breasts were normal. The CMMD includes 2601 breasts (1775 patients) from multiple centers, 556 breast biopsies were confirmed benign, 1316 breast biopsies were confirmed malignant, and 729 breasts were normal. Each breast in both open-source datasets has paired images for both CC and MLO views, resulting in a total of 7808 mammogram images for the Mini-DDSM dataset and 5202 mammogram images for the CMMD dataset. Figure 2 and Figure 3 shows examples of four views of a patient's left and right breast in the Mini-DDSM dataset and CMMD dataset, respectively. In this study, the mammogram images from the above two open-source datasets were classified, with normal and benign breasts considered positive cases and malignant breasts considered negative cases. We divided each dataset into a training set and a test set in the ratio of 80:20, and the results of the division of the number of images and the number of breasts in each dataset are given in Table 1.

The Mini-DDSM used in this study was derived from kaggle's processed Mini-DDSM's 16-bit PNG (portable network graphics) images. For the CMMD dataset, we converted all raw DICOM images to lossless 8-bit JPEG (joint photographic experts group) images for subsequent processing.

TABLE 1. Summary of the number of images and the number of breasts in each dataset and subset.

Dataset		Train（80%）		Test（20%）		Total
		Normal/ Benign	Mali gnant	Normal/ Benign	Malig nant	
Mini-DDSM	Image number	5052	1192	1276	288	7808
	Breast number	2526	596	638	144	3904
CMMD	Image number	2060	2104	510	528	5202
	Breast number	1030	1052	255	264	2601

B. DATA PRE-PROCESSING

This section provides a detailed introduction to the data pre-processing process and the required visualization.

1) BREAST REGION SEGMENTATION

The size of the open source dataset Mini-DDSM is (495-2746) × (1088-3481) and the size of CMMD is 1914 × 2294. Although the higher resolution of mammogram images contains more information, training deep learning models with high-resolution original images, the following challenges still exist:

1) Since the original mammogram images are too large, direct resizing may lose information about some lesions, leading to models that are unable to learn from these lesions [22].

2) As shown in Figure 2, the Mini-DDSM dataset's original mammogram images frequently contain undesirable view label information, which will certainly reduce the classification accuracy.

3) As shown in Figure 3, the CMMD dataset's original mammogram images contain a significant amount of redundant regions. The lesions are only present in the mammary region, which still accounts for less than half of the mammogram images. In addition to being useless for categorizing lesions as benign or malignant, redundant regions also interfere with model training and raise computing costs [23].

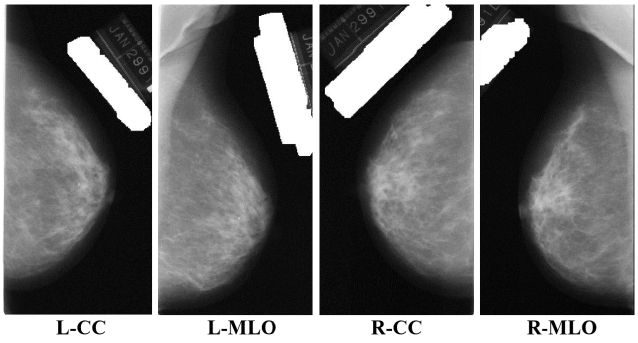

FIGURE 2. Four views for the right and left breast mammography images from the public datasets of Mini-DDSM.

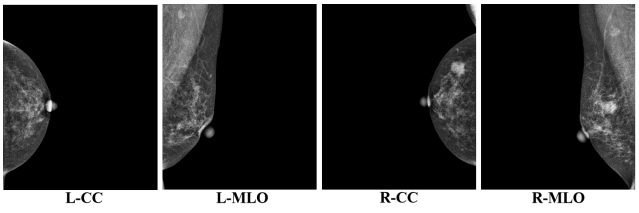

FIGURE 3. Four views for the right and left breast mammography images from the public datasets of CMMD.

To solve the above problem, we use the BRS (breast region segmentation) module to pre-process the original mammogram images. Figure 4 shows the processing process of this module on the Mini-DDSM dataset. The module first replaces the pixel values larger than 254 with 0, and then detects the edges of the tissue to obtain the coordinates of the four corners of the ROI (region of interest) box. Based on the coordinates of the four corners of the ROI box, the breast region is obtained by cropping on the original image. Finally, the mammary region is resized to a fixed size of 640 × 640 pixels and used as the input to the model. The module processes the CMMD dataset similarly to the Mini-DDSM dataset, the

only difference is that the CMMD dataset does not have white areas and does not need to replace pixel values larger than 254 to eliminate the effect of white areas on the detection of breast tissue.

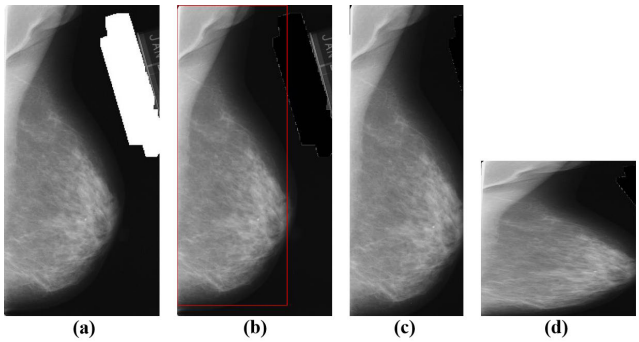

FIGURE 4. Processing of the BRS module on the Mini-DDSM dataset. (a) The original image; (b) ROI rectangle; (c) The cropped image; (d) The resized image.

2) DATA AUGMENTATION

To improve the robustness and generalization of the model, we enhanced the training set images with four data augmentation methods, all with probability values set to 0.5. The four data augmentation methods were a) flipping the images in the vertical direction; b) flipping the images in the horizontal direction; c) affine transformation with a rotate value of 20, a translate_percent value of 0.1, a shear value of 20, and a scale value of 0.8 to 1.2; d) elastic transformation with an alpha value of 10 and a sigma value of 15. And all images of the dataset were normalized. Each training image in the final dataset is eight times larger than it was before augmentation. Following augmentation, 49,952 training images of Mini-DDSM and 33,312 training images of CMMD were obtained. Figure 5 depicts an example of enhanced mammography images for these four data enhancement methods.

C. PROPOSED LCVT-GR OVERALL ARCHITECTURE

The overall architecture of LCVT-GR is shown in Figure 6. The input of the model is the images of two views of the breast (CC view and MLO view). The input images are first extracted from the features U_{CC} and U_{MLO} of these two views by a backbone model, here the backbone model used in this study is tf_efficientnetv2_s. Then, the features extracted from the backbone model are passed through the Local Cross-View Transformers Module (LCVTM) and Global Representation Module (GRM) in parallel to generate the local and global representations. Finally, the local and global representations are concatenated into the MLP (Multi-Layer Perception) classification layer to generate the prediction results.

D. LOCAL CROSS-VIEW TRANSFORMERS MODULE

LCVTM models the local semantic relationship between two views to generate a local representation. To better learn the dependencies between the two views, LCVTM uses the idea

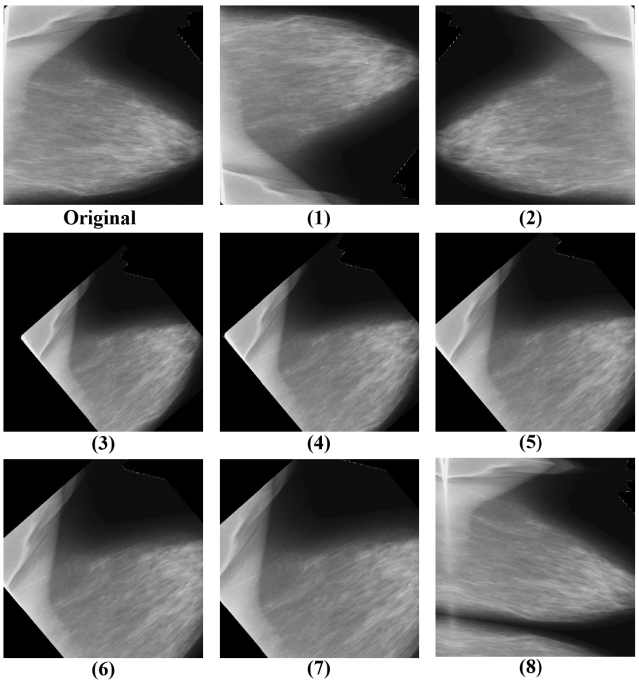

FIGURE 5. Sample augmented images from different augmentation methods. (1–2): Image flipping in both vertical and horizontal direction, (3–7): Affine transformation with a value of scale from 0.8 to 1.2, (8): Elastic transformation.

of cross transformers. Where the attention mechanism of the transformer follows the cross-shaped window self-attention method in CSWin [24]. Since our model LCVT-GR inputs the information of two views, in order to achieve the interaction of U_{CC} and U_{MLO} information, we design a Cross-View Attention Module (CVAM).

1) CROSS-SHAPED WINDOW SELF-ATTENTION

Transformer uses a self-attention mechanism to model contextual information in order to capture long-range dependencies. However, this pixel point pair-based modeling approach necessitates a significant amount of computation, often the quadratic power of the input feature size [25]. Therefore, the computational cost consumption can be very high when the input feature map resolution is relatively high. Swin et al. [26] recommended the usage of local windows self-attention to broaden the field of perception through shift windows in order to solve this issue. However, this still does not address the issue of the token's constrained attention area within the transformer block. To expand the attention area more effectively, CSWin [24] proposed the cross-shaped window self-attention mechanism, which implements self-attention by forming horizontal and vertical stripes of the cross-shaped window, which results in a wider receptive field for the token within each transformer blocks with stronger contextual modeling capabilities.

The schematic diagram of cross-shaped window self-attention is shown in Figure 7. Cross-shaped window self-attention is based on a multi-head self-attention

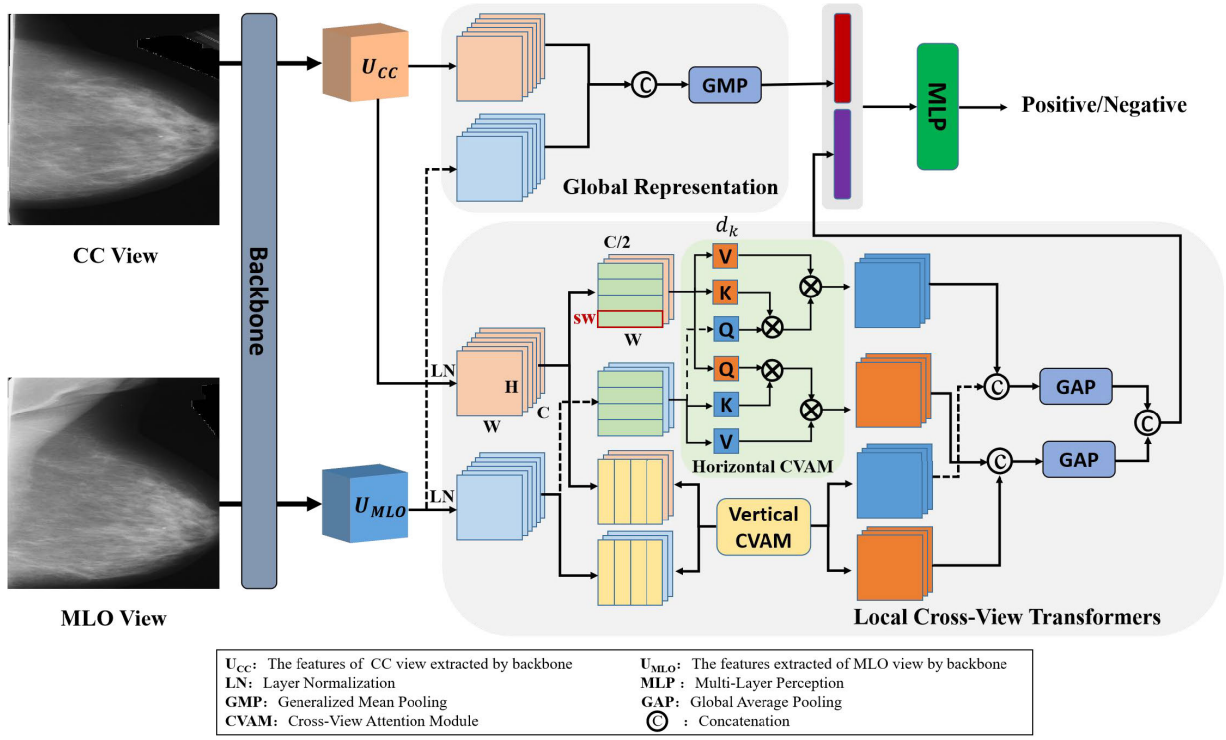

FIGURE 6. The overall architecture of our proposed model LCVT-GR.

mechanism by first linearly projecting the input feature $X \in R^{(H\times W)\times C}$ onto K heads ($\{h_1, \ldots, h_K\}$), and then dividing the K heads equally into two parallel groups, each with the number of channels C/2. The two groups of heads apply different ways of self-attention, with the first group of heads performing horizontal striped self-attention and the second group of heads performing vertical striped self-attention, both in parallel. Finally, the outputs of these two groups are connected. Unlike the cross-shaped window self-attention method, we utilize the CVAM to perform cross-attention on the information of the two views (U_{CC} and U_{MLO}) of the input LCVTM. The horizontal CVAM exchanges the Q generated by the first group of U_{CC} headers and the first group of U_{MLO} headers, and the vertical CVAM exchanges the Q generated by the second group of U_{CC} headers and the second group of U_{MLO} headers. CVAM implements cross-attention of two views by exchanging the Q generated by horizontal stripe self-attention and the Q generated by vertical stripe self-attention of both views. CVAM achieves cross-attention by referring to the local co-occurrence module proposed in the paper [27], both by exchanging Qs generated by different features to achieve information interaction. Assuming that after CVAM, the output of U_{CC} is A and the output of U_{MLO} is B, the output of LCVTM can be defined as:

$$\mathrm{LCVTM}\left(U_{CC}, U_{MLO}\right) = \mathrm{Concat}\left(\mathrm{GAP}\left(A\right), \mathrm{GAP}\left(B\right)\right) \quad (1)$$

$$A = \mathrm{Concat}(A_1, \ldots, A_k, \ldots, A_K)W^O k = 1, \ldots, K \quad (2)$$

$$B = \mathrm{Concat}(B_1, \ldots, B_k, \ldots, B_K)W^O k = 1, \ldots, K \quad (3)$$

$W^O \in R^{C\times C}$ denotes the projection matrix and the output dimension is set to C. GAP stands for global average pooling and LN stands for layer normalization.

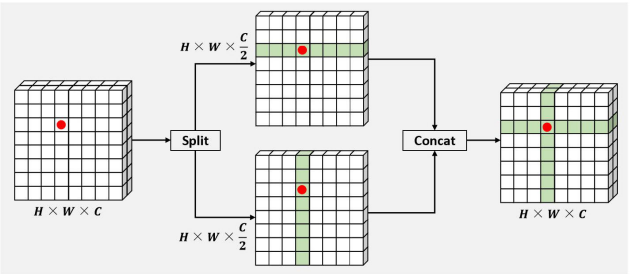

FIGURE 7. The diagram of cross-shaped window self-attention.

2) CROSS-VIEW ATTENTION MODULE

For the cross attention of both U_{CC} and U_{MLO} views, U_{CC} is uniformly divided into non-overlapping equal high horizontal stripes $\left[U^1_{CC}, \ldots, U^m_{CC}, \ldots, U^M_{CC}\right]$ and non-overlapping equal wide vertical stripes $\left[U^1_{CC}, \ldots, U^z_{CC}, \ldots, U^Z_{CC}\right]$, and U_{MLO} performs the same operation. Sw is the dynamic stripe width when dividing the stripes, and sw can be adjusted to balance the learning ability and computational complexity of the model.

$$\left[U^1_{CC}, \ldots, U^m_{CC}, \ldots, U^M_{CC}\right] = U_{CC}, U^m_{CC} \in R^{(sw\times W)\times C}, M = H/sw \quad (4)$$

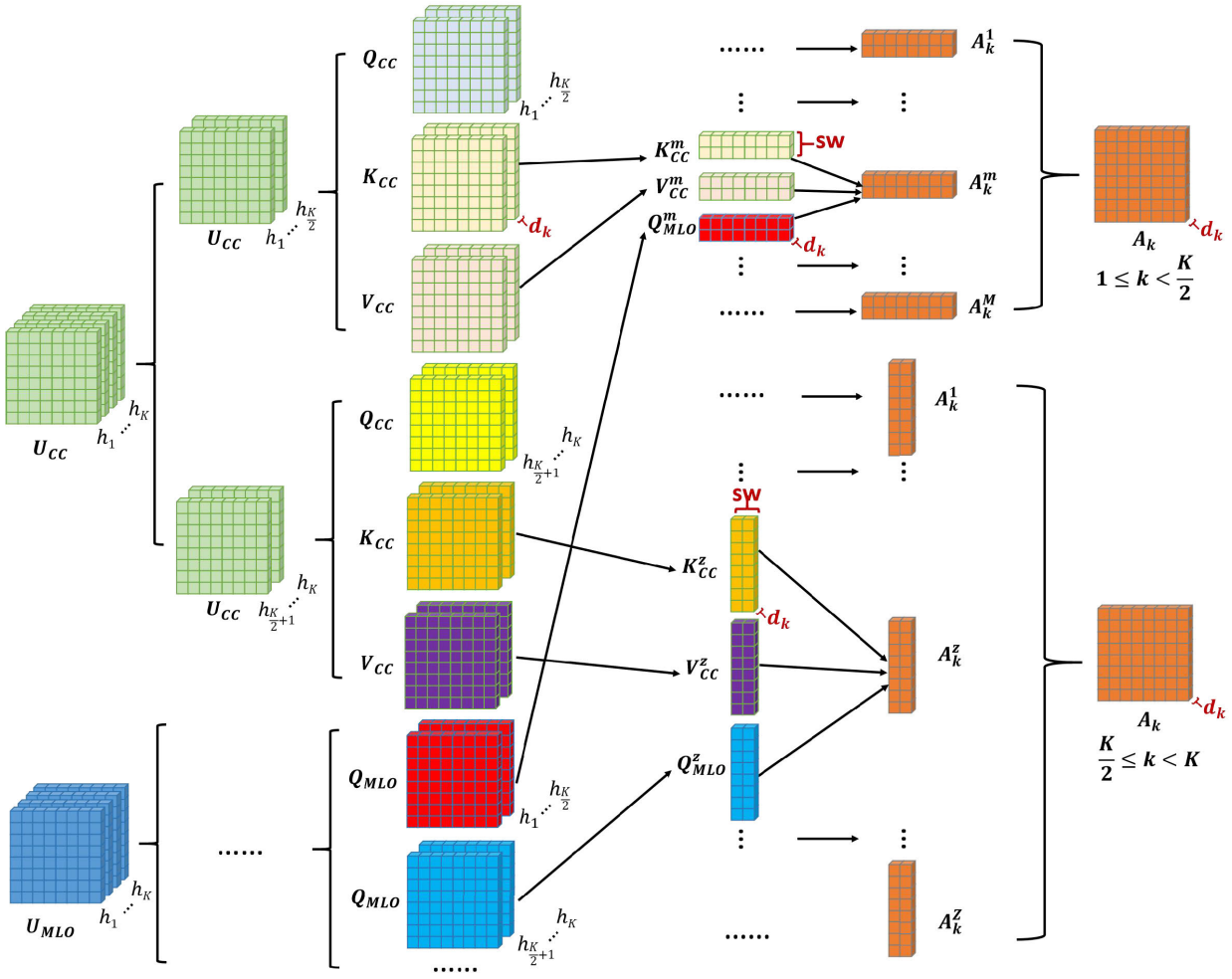

FIGURE 8. The diagram for calculating A_k.

$$\left[U^1_{CC},\ldots,U^z_{CC},\ldots,U^Z_{CC}\right] = U_{CC}, U^z_{CC}\in R^{(sw\times H)\times C}, Z = W/sw \tag{5}$$

$$\left[U^1_{MLO},\ldots,U^m_{MLO},\ldots,U^M_{MLO}\right] = U_{MLO}, U^m_{MLO}\in R^{(sw\times W)\times C}, M = H/sw \tag{6}$$

$$\left[U^1_{MLO},\ldots,U^z_{MLO},\ldots,U^Z_{MLO}\right] = U_{MLO}, U^z_{MLO}\in R^{(sw\times H)\times C}, Z = W/sw \tag{7}$$

Assuming that the projection Q(query), K(key) and V(value) dimensions of the k^{th} head are d_k, the output of the k^{th} head of U_{CC} after cross attention is defined as:

$$A_k = \begin{cases} H-CAttn_k\,(U_{CC}, U_{MLO}) \\ = \left[A^1_k,\ldots,A^m_k,\ldots,A^M_k\right] \quad k = 1,\ldots,K/2 \\ V-CAttn_k\,(U_{CC}, U_{MLO}) \\ = \left[A^1_k,\ldots,A^z_k,\ldots,A^Z_k\right] \quad k = \frac{K}{2}+1,\ldots,K \end{cases} \tag{8}$$

$$A^m_k = CVAM(Q^m_{MLO}, K^m_{CC}, V^m_{CC}) \tag{9}$$

$$A^z_k = CVAM(Q^z_{MLO}, K^z_{CC}, V^z_{CC}) \tag{10}$$

$$Q^m_{MLO} = U^m_{MLO}W^Q_k, K^m_{CC} = U^m_{CC}W^K_k, V^m_{CC} = U^m_{CC}W^V_k \tag{11}$$

$$Q^z_{MLO} = U^z_{MLO}W^Q_k, K^z_{CC} = U^z_{CC}W^K_k, V^z_{CC} = U^z_{CC}W^V_k \tag{12}$$

$W^Q_k \in R^{C\times d_k}, W^K_k \in R^{C\times d_k}, W^V_k \in R^{C\times d_k}$ denote the projection matrix of Q, K and V of the k^{th} head, respectively. d_k denotes the channel dimension of the kth head with the value C/K. The output of cross-attention of horizontal stripes to the k^{th} head is noted as $H-CAttn_k(X_1, X_2)$, and the output of cross-attention of vertical stripes to the k^{th} head is noted as $V-CAttn_k(X_1, X_2)$. The Q and K, V of CVAM come from the features of different views, respectively, and the CVAM is calculated as:

$$CVAM\left(Q^m_{MLO}, K^m_{CC}, V^m_{CC}\right) = \text{softmax}\left(\frac{Q^m_{MLO}\left(K^m_{CC}\right)^T}{\sqrt{d_k}}\right)V^m_{CC} \tag{13}$$

TABLE 2. Testing results show breast-level estimates of AUC-ROC and AUC-RP on Mini-DDSM and CMMD.

Datasets	Models	View approach	#Params(M)	Batch size	Testing AUC-ROC	Testing AUC-PR
Mini-DDSM	Efficientnet-B2	Single-view	7.7	16	0.8474(0.8223-0.8698)	0.6286(0.5750-0.6838)
	Resnet101	Single-view	42.5	16	0.8060(0.7793-0.8333)	0.5187(0.4584-0.5839)
	Densenet121	Single-view	6.9	16	0.8140(0.7878-0.8408)	0.5369(0.4748-0.5979)
	Mobilenetv3-large	Single-view	4.2	56	0.8276(0.8027-0.8515)	0.5671(0.5091-0.6249)
	PHResNet18	Multi-view	5.6	4	0.6768(0.6321-0.7215)	0.3038(0.2431-0.3773)
	Breast-wide-model	Multi-view	3.1	216	0.6852(0.6425-0.7312)	0.2789(0.2290-0.3374)
	Two-views-classifier	Multi-view	86.2	8	0.7056(0.6626-0.7480)	0.3272(0.2643-0.4039)
	LCVT-GR	**Multi-view**	**36.2**	**8**	**0.8585(0.8250-0.8899)**	**0.6576(0.5749-0.7297)**
CMMD	Efficientnet-B2	Single-view	7.7	16	0.8442(0.8192-0.8670)	0.8714(0.8464-0.8935)
	Resnet101	Single-view	42.5	16	0.8577(0.8336-0.8794)	0.8740(0.8443-0.8988)
	Densenet121	Single-view	6.9	16	0.8321(0.8070-0.8555)	0.8467(0.8136-0.8765)
	Mobilenetv3-large	Single-view	4.2	56	0.8558(0.8320-0.8781)	0.8770(0.8526-0.8995)
	PHResNet18	Multi-view	5.6	4	0.7141(0.6680-0.7596)	0.7164(0.6542-0.7767)
	Breast-wide-model	Multi-view	3.1	216	0.7657(0.7248-0.8057)	0.7866(0.7335-0.8405)
	Two-views-classifier	Multi-view	86.2	8	0.7679(0.7269-0.8071)	0.7709(0.7139-0.8282)
	LCVT-GR	**Multi-view**	**36.2**	**8**	**0.8712(0.8410-0.9013)**	**0.8903(0.8598-0.9190)**

Figure 8 shows how the output A_k of the k^{th} head of U_{CC} after cross-attention is obtained, and the output of the k^{th} head of U_{MLO} after cross-attention can be derived similarly.

3) GLOBAL REPRESENTATION MODULE

Although LCVTM adopts CSWin's cross-shaped window self-attention method, which effectively expands the attention region, LCVTM lacks the global information of the image. Therefore, we designed the GRM component to extract the global information of the image. GRM combines the features U_{CC} and U_{MLO} of the two views extracted by backbone and performs feature dimensionality reduction by GMP (generalized mean pooling).

$$\mathrm{GRM}\left(U_{CC}, U_{MLO}\right) = \mathrm{GMP}\left(\mathrm{Concat}(U_{CC}, U_{MLO})\right) \quad (14)$$

The global representation information generated by GRM, which is extracted from the whole image, can compensate for LCVTM's lack of global information extraction. The efficiency of the GRM component will be demonstrated in the ablation experiment section.

III. RESULTS AND DISCUSSION

A. IMPLEMENTATION DETAILS

We use AdamW optimizer [28] to train the models to minimize the binary cross-entropy (BCE) loss with a learning rate of 0.0001 and a weight decay of 0.01. The batch size is 8 for the multi-view model and 16 for the single-view model, and all models are trained for 20 epoch. We use OneCycleLR to dynamically control the learning rate reduction based on BCE loss during model training, where the maximum learning rate is 0.0001 and the proportion of the learning rate increase is 0.1. All experiments are implemented using PyTorch and performed on an NVIDIA GTX 1080 Ti GPU (12GB).

B. EVALUATION METRICS

We refer to the paper [14], [27] to evaluate the classification results based on two metrics: the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision-recall curve (AUC-PR). The AUC-ROC and AUC-PR are common metrics used to assess the performance of radiologists, enabling the assessment of model performance and the comparison of differences between models. The AUC-ROC reflects the balance between the model's TPR (true positive rate also called recall) and FPR (false positive rate) at different probability thresholds. The higher the AUC-ROC value of a model, the better its ability to distinguish between positive and negative cases. The TPR and FPR are evaluated as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \quad (15)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{(\mathrm{FP} + \mathrm{TN})} \quad (16)$$

where the letters TP, FN, FP, and TN stand for the corresponding totals of true positive, false negative, false positive, and true negative samples. The AUC-PR reflects the balance between the model's recall and precision (positive predictive value) at different probability thresholds. The precision is defined as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \quad (17)$$

C. RESULTS AND DISCUSSIONS

Table 2 shows the results of comparing our model with several two-view mammogram image classification models as well as traditional CNN classification models on the test datasets of Mini-DDSM and CMMD. When training competing models, our batch size is no longer 8 or 16, but is set as large as possible to make full use of the GPU memory. In Table 2, we give the values of the batch size settings for the model during training. PHResNet18 [10] is a two-view breast cancer classification method based on parameterized hypercomplex neural networks proposed by Lopez et al. PHResNet18 uses ResNet18 as a backbone to model the

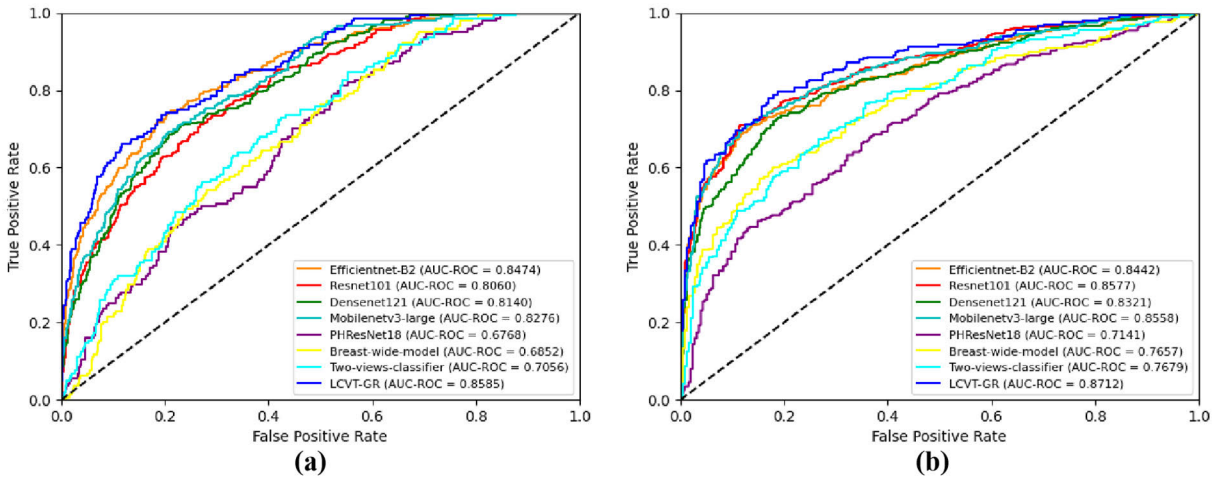

FIGURE 9. The ROC curves of comparison results of different models on different datasets. (a) Mini-DDSM; (b) CMMD.

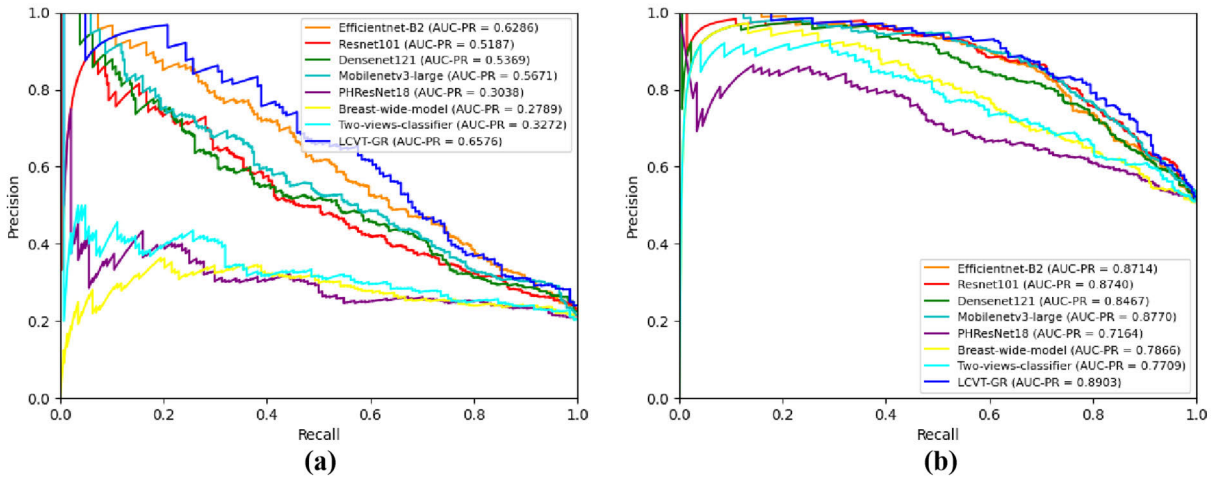

FIGURE 10. The PR curves of comparison results of different models on different datasets. (a) Mini-DDSM; (b) CMMD.

correlations that exist between different views using hypercomplex algebraic properties. The breast-wide-model [14] is a two-branch, two-view breast cancer classification model with ResNet22 as the backbone, as proposed by Wu et al. The model extracts features from different views for each branch and finally aggregates the extracted features for the final prediction. Two-views-classifier [13] is a two-view breast cancer classification model based on three-time migration learning proposed by Petrini et al. The model first trains the patched classifier on natural images, then trains the one-view classifier using the patched classifier weights, and finally trains the two-view classifier using the one-view classifier weights. Since in the experimental part we only compare the network structure of the two-views-classifier, the migration learning method mentioned in the paper is not used.

We used bootstrap resampling (2000 bootstrap repetitions) to estimate the 95% CI (confidence interval) in our tests and give the mean, lower, and upper values of the 95% CI for both AUC-ROC and AUC-PR metrics in Table 2. In addition, as shown in Fig. 9 and Fig. 10, we plotted the ROC curves and PR curves of this paper's model and the competing models on the Mini-DDSM and CMMD datasets, respectively, to visualize the classification performance. From the test results, we can see that this paper's model significantly outperforms the competing models. On the Mini-DDSM test dataset, AUC-ROC reaches 85.85% and AUC-PR reaches 65.76%, with an average improvement of 9.24% in AUC-ROC and 20.6% in AUC-PR. On the CMMD test dataset, AUC-ROC reached 87.12% and AUC-PR reached 89.03%, with an average improvement of 6.58% in AUC-ROC and 6.99% in AUC-PR.

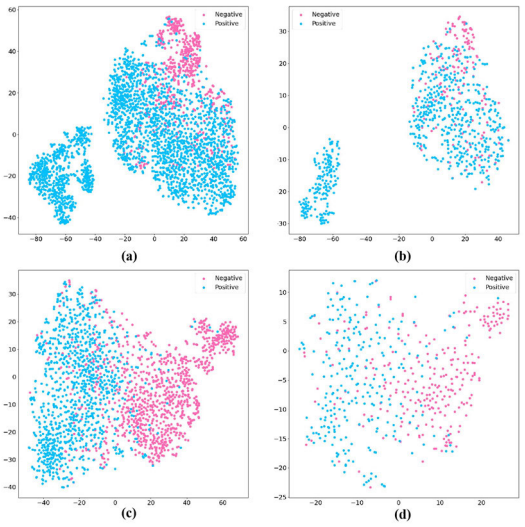

FIGURE 11. TSNE plots of the (a) Mini-DDSM training set; (b) Mini-DDSM test set; (c) CMMD training set; (d) CMMD test set.

The variability of features can be assessed qualitatively using TSNE (t-distributed stochastic neighbor embedding) [29] plots. For each pair of training and test samples from the Mini-DDSM and CMMD datasets, we plotted the features from both local and global analysis using TSNE plots. As shown in Fig. 11, our model can extract discriminative features with discrepancies for classification.

D. ABLATION STUDIES

By comparing LCVT-GR with several two-view mammogram image classification models as well as traditional CNN classification models, we demonstrate that LCVT-GR has the best classification performance, which likewise proves that our strategy of using the global-local parallel analysis method to mimic the manual analysis is effective. Besides, to validate the soundness of the LCVT-GR design, we also discuss the structure of LCVT-GR, for which we perform three ablation experiments.

1) KEY COMPONENTS

To evaluate the effectiveness of the model components, we conducted ablation experiments on each component of LCVT-GR, and Table 3 shows the results of the ablation experiments on both Mini-DDSM and CMMD datasets. LTM indicates that in the LCVTM module, no interaction between the two views is performed. On the Mini-DDSM dataset, single-view classification prediction using the backbone network (tf_efficientnetv2_s) achieves 84.04% AUC-ROC and 61.5% AUC-PR (shown in the first row of Table 3).

As can be demonstrated in the second and third rows of Table 3, the AUC-ROC or AUC-PR of the test results is improved using Backbone+LCVTM or Backbone+GRM for multi-view classification prediction, indicating that the LCVTM and GRM components are effective. Thus, when Backbone+LCVTM+GRM (LCVT-GR) is used for multi-view classification prediction, the boosting effect of the two components adds up to an increase in AUC-ROC to 85.85% and AUC-PR to 65.76% (shown in the sixth row of Table 3). In addition to this, comparing the fourth, fifth, and sixth rows of Table 3 shows that making the two views cross-attention when extracting local representation information improves the AUC-ROC by 1% and the AUC-PR by 1.05%. Using the multi-view classification model gives better predictions than using the single-view classification model, with a maximum improvement of 1.99% in AUC-ROC and 3.09% in AUC-PR. Looking at the results of the ablation experiments for both the Mini-DDSM and CMMD datasets, we found that they have the same pattern. Therefore, we conclude that all components of LCVT-GR are effective and that the structure of LCVT-GR is optimal.

2) DYNAMIC STRIPE WIDTH

In Table 4, we examine the trade-off between stripe width (sw) and model performance on the Mini-DDSM dataset. We find that the computational costs (FLOPs) increase as the stripe width increases, but the performance of the model does not increase with it. We come to the conclusion that when sw is set to 1, better model performance can be attained with the least amount of computational expense.

3) NUMBER OF HEADS

In Table 5, we also examine the trade-off between the number of heads (K) and the model performance on the Mini-DDSM dataset. We find that the performance of the model improves significantly at the beginning as K increases and decreases when K is large enough. The value of K affects the performance of the model but does not change the computational costs (FLOPs). We believe that when sw is set to 1, K is set to 4, which leads to the optimal performance of the model. Based on the results of this ablation experiment, the LCVT-GR model for all experiments in this paper is set to sw $= 1$ and K $= 4$ by default.

In summary, the first ablation experiment demonstrates that a) the key components of LCVT-GR are all valid, b) the classification results are better with two views of information interacting than without, and c) the structure of two views performs better than the structure of a single view. This is in line with our initial assumptions. The images of two views of a breast often contain complementary information, so the classification performance of the two-view model is better than that of the single-view model. The information interaction between the two views can learn more dependencies, which is important for improving the detection rate of lesions. In the second and third experiments, we ablated two variables of the model, sw and K, respectively, and finally determined that the model achieves a good trade-off between computational cost and performance when sw $= 1$ and K $= 4$.

TABLE 3. Ablation study of key components of our LCVT-GR model.

Datasets	Backbone	LTM	LCVTM	GRM	View approach	Testing AUC-ROC	Testing AUC-PR
Mini-DDSM	√				Single-view	0.8404(0.8145-0.8660)	0.6150(0.5594-0.6684)
	√		√		Multi-view	0.8393(0.8056-0.8730)	0.6323(0.5534-0.7042)
	√			√	Multi-view	0.8470(0.8143-0.8799)	0.6326(0.5529-0.7057)
	√	√		√	Multi-view	0.8485(0.8133-0.8807)	0.6471(0.5660-0.7225)
	√	√		√	Single-view	0.8386(0.8144-0.8625)	0.6267(0.5720-0.6823)
	√		√	√	**Multi-view**	**0.8585(0.8250-0.8899)**	**0.6576(0.5749-0.7297)**
CMMD	√				Single-view	0.8369(0.8115-0.8602)	0.8644(0.8375-0.8873)
	√		√		Multi-view	0.8547(0.8225-0.8856)	0.8826(0.8514-0.9101)
	√			√	Multi-view	0.8682(0.8347-0.8976)	0.8890(0.8562-0.9184)
	√	√		√	Multi-view	0.8661(0.8338-0.8970)	0.8832(0.8483-0.9151)
	√	√		√	Single-view	0.8456(0.8231-0.8687)	0.8690(0.8426-0.8931)
	√		√	√	**Multi-view**	**0.8712(0.8410-0.9013)**	**0.8903(0.8598-0.9190)**

TABLE 4. Ablation on stripe width (sw) in the Mini-DDSM dataset.

sw	K	#Params(M)	FLOPs(G)	Testing AUC-ROC	Testing AUC-PR
1	**4**	**36.2**	**51.36**	**0.8585(0.8250-0.8899)**	**0.6576(0.5749-0.7297)**
2	4	36.2	51.40	0.8415(0.8072-0.8738)	0.6204(0.5421-0.6953)
4	4	36.2	51.48	0.8552(0.8225-0.8875)	0.6606(0.5865-0.7290)
5	4	36.2	51.52	0.8399(0.8042-0.8734)	0.6484(0.5712-0.7167)
10	4	36.2	51.72	0.8399(0.8060-0.8741)	0.6246(0.5472-0.6991)

TABLE 5. Ablation on the number of heads in Mini-DDSM dataset.

sw	K	Testing AUC-ROC	Testing AUC-PR
1	2	0.8413(0.8075-0.8726)	0.6304(0.5554-0.7017)
1	**4**	**0.8585(0.8250-0.8899)**	**0.6576(0.5749-0.7297)**
1	8	0.8479(0.8143-0.8800)	0.6352(0.5529-0.7097)
1	16	0.8437(0.8079-0.8759)	0.6333(0.5529-0.7056)

With these three ablation experiments, we demonstrate that the structure of LCVT-GR is scientific and effective.

IV. CONCLUSION

In this paper, we propose a new multi-view mammography image classification method that uses a two-view global-local parallel analysis method to extract global and local information about mammography images. Global analysis helps to detect abnormalities such as distortion and asymmetric denseness of breast structures, and local analysis helps to detect abnormalities such as masses and calcifications. In order to better learn the dependencies between two views and realize the information exchange between different view features, we employ the cross-transformer concept while extracting local information.

To validate the effectiveness of our method, we conducted comparison experiments and ablation experiments on two publicly available datasets, Mini-DDSM and CMMD. The results of the comparison experiments show that our method achieves better results compared with existing advanced methods, with greater improvements in both AUC-ROC and AUC-PR assessment metrics. The results of the ablation experiments show that our model architecture is scientific and effective and achieves a good trade-off between computational cost and model performance.

Our proposed method will help to build high-performance and robust deep learning-based mammography CAD systems to improve efficiency and reduce cost in the early diagnosis of breast cancer. Since the public dataset used in this study experiment is still relatively small, we will further validate our model on a larger dataset in future work.

CONFLICTS OF INTEREST

None declared.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] I. Hoxha, D. A. Islami, G. Uwizeye, V. Forbes, and M. D. Chamberlin, "Forty-five years of research and progress in breast cancer: Progress for some, disparities for most," *JCO Global Oncol.*, vol. 8, Apr. 2022, Art. no. e2100424.

[3] D. Barba, A. León-Sosa, P. Lugo, D. Suquillo, F. Torres, F. Surre, L. Trojman, and A. Caicedo, "Breast cancer, screening and diagnostic tools: All you need to know," *Crit. Rev. Oncol./Hematol.*, vol. 157, Jan. 2021, Art. no. 103174.

[4] Canadian Task Force on Preventive Health Care, "Recommendations on screening for breast cancer in average-risk women aged 40–74 years," *Cmaj*, vol. 183, no. 17, pp. 1991–2001, 2011.

[5] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. Saripan, A. R. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review," *Clin. Imag.*, vol. 37, no. 3, pp. 420–426, May/Jun. 2013.

[6] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Academic Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.

[7] Y. Guo, Y. Cai, Z. Cai, Y. Gao, N. An, L. Ma, S. Mahankali, and J. Gao, "Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging," *J. Magn. Reson. Imag.*, vol. 16, no. 2, pp. 172–178, Aug. 2002.
[8] M. Samulski and N. Karssemeijer, "Optimizing case-based detection performance in a multiview CAD system for mammography," *IEEE Trans. Med. Imag.*, vol. 30, no. 4, pp. 1001–1009, Apr. 2011.
[9] M. Velikova, M. Samulski, P. J. F. Lucas, and N. Karssemeijer, "Improved mammographic CAD performance using multi-view information: A Bayesian network framework," *Phys. Med. Biol.*, vol. 54, no. 5, pp. 1131–1147, Mar. 2009.
[10] E. Lopez, E. Grassucci, M. Valleriani, and D. Comminiello, "Multi-view hypercomplex learning for breast cancer screening," 2022, *arXiv:2204.05798*.
[11] R. Rashmi, K. Prasad, and C. B. K. Udupa, "BCHisto-Net: Breast histopathological image classification by global and local feature aggregation," *Artif. Intell. Med.*, vol. 121, Nov. 2021, Art. no. 102191.
[12] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci. Rep.*, vol. 9, no. 1, p. 12495, Aug. 2019.
[13] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, "Breast cancer diagnosis in two-view mammography using end-to-end trained EfficientNet-based convolutional network," *IEEE Access*, vol. 10, pp. 77723–77731, 2022.
[14] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, and S. Wolfson, "Deep neural networks improve Radiologists' performance in breast cancer screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020.
[15] X. Chen, K. Zhang, N. Abdoli, P. W. Gilley, X. Wang, H. Liu, B. Zheng, and Y. Qiu, "Transformers improve breast cancer diagnosis from unregistered multi-view mammograms," *Diagnostics*, vol. 12, no. 7, p. 1549, Jun. 2022.
[16] K. Wang, B. He, and W.-P. Zhu, "Cptnn: Cross-parallel transformer neural network for time-domain speech enhancement," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2022, pp. 1–5.
[17] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5321–5330.
[18] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowl.-Based Syst.*, vol. 264, Mar. 2023, Art. no. 110362.
[19] Y. Li, Y. Sun, Z. Cui, S. Shan, and J. Yang, "Learning fair face representation with progressive cross transformer," 2021, *arXiv:2108.04983*.
[20] C. D. Lekamlage, F. Afzal, E. Westerberg, and A. Cheddad, "Mini-DDSM: Mammography-based automatic age estimation," in *Proc. 3rd Int. Conf. Digit. Med. Image Process.*, 2020, pp. 1–6.
[21] C. Cui, L. Li, H. Cai, Z. Fan, L. Zhang, T. Dan, J. Li, and J. Wang, "The Chinese Mammography Database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast," *Cancer Imag. Arch.*, 2021, doi: 10.7937/tcia.eqde-4b16.
[22] L. Xie, L. Zhang, T. Hu, H. Huang, and Z. Yi, "Neural networks model based on an automated multi-scale method for mammogram classification," *Knowl.-Based Syst.*, vol. 208, Nov. 2020, Art. no. 106465.
[23] D. Muduli, R. Dash, and B. Majhi, "Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 102825.
[24] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12124–12134.
[25] Z. Shen, I. Bello, R. Vemulapalli, X. Jia, and C.-H. Chen, "Global self-attention networks for image recognition," 2020, *arXiv:2010.03019*.
[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[27] Y. Chen, H. Wang, C. Wang, Y. Tian, F. Liu, Y. Liu, M. Elliott, D. J. McCarthy, H. Frazer, and G. Carneiro, "Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation," in *Proc. 25th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Singapore, Cham, Switzerland: Springer, 2022, pp. 3–13.
[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[29] S. Dey, D. Mandal, L. D. Robertson, B. Banerjee, V. Kumar, H. McNairn, A. Bhattacharya, and Y. S. Rao, "In-season crop classification using elements of the Kennaugh matrix derived from polarimetric RADARSAT-2 SAR data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 88, Jun. 2020, Art. no. 102059.

WENNA WU received the B.M. degree from Changsha University of Science and Technology, in 2021. She is currently pursuing the master's degree with the School of Biomedical Engineering, Southern Medical University. Her research interests include radiomics and medical image processing.

QI RONG received the B.M.E. degree from the University of South China, in 2022. She is currently pursuing the master's degree with the School of Biomedical Engineering, Southern Medical University. Her research interests include radiomics and medical image processing.

ZHENTAI LU (Member, IEEE) received the M.S. degree in mathematics from Sun Yat-sen University, Guangzhou, China, and the Ph.D. degree from the School of Biomedical Engineering, Southern Medical University, Guangzhou, in 2008. He is currently a Professor with the School of Biomedical Engineering, Southern Medical University. His research interests include pattern recognition, machine learning, and image processing.