Dual-view Correlation Hybrid Attention Network for Robust Holistic Mammogram Classification

Zhiwei Wang[1,2], **Junlin Xian**[3], **Kangyi Liu**[3], **Xin Li**[1,2], **Qiang Li**[1,2] and **Xin Yang**[3]
[1] Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
[2] MoE Key Laboratory for Biomedical Photonics, Collaborative Innovation Center for Biomedical Engineering, School of Engineering Sciences, Huazhong University of Science and Technology
[3] School of Electronic Information and Communications, Huazhong University of Science and Technology
{zwwang, xianjunlin, m20202099, lixin_bme, liqiang8, xinyang2014}@hust.edu.cn

Abstract

Mammogram image is important for breast cancer screening, and typically obtained in a dual-view form, i.e., cranio-caudal (CC) and mediolateral oblique (MLO), to provide complementary information. However, previous methods mostly learn features from the two views independently, which violates the clinical knowledge and ignores the importance of dual-view correlation. In this paper, we propose a dual-view correlation hybrid attention network (DCHA-Net) for robust holistic mammogram classification. Specifically, DCHA-Net is carefully designed to extract and reinvent deep features for the two views, and meanwhile to maximize the underlying correlations between them. A hybrid attention module, consisting of local relation and non-local attention blocks, is proposed to alleviate the spatial misalignment of the paired views in the correlation maximization. A dual-view correlation loss is introduced to maximize the feature similarity between corresponding strip-like regions with equal distance to the chest wall, motivated by the fact that their features represent the same breast tissues, and thus should be highly-correlated. Experimental results on two public datasets, i.e., INbreast and CBIS-DDSM, demonstrate that DCHA-Net can well preserve and maximize feature correlations across views, and thus outperforms the state-of-the-arts for classifying a whole mammogram as malignant or not.

1 Introduction

Breast cancer is the most common malignant tumor of middle-aged and elderly women, and around 1.2 million women are diagnosed with breast cancer every year [Hosseini *et al.*, 2016]. The classification of breast cancer mainly relies on the immunohistochemical diagnosis of breast cancer tissue, which is complex and traumatic and thus can not meet the needs of accurate diagnosis and personalized treatment. With the development of medical imaging techniques, radiomics-based breast cancer diagnosis [Lu *et al.*, 2018] has become a new non-invasive cancer assessment approach, which is comprehensive, easy to obtain and economic.

Among the imaging modalities, mammography has been proven to be effective in early detection and diagnosis [Moss *et al.*, 2012]. In a standard mammographic screening examination, the 3D breast will be projected onto a 2D X-ray film. Typically, each breast will be exposure in two different angles, i.e., cranio-caudal (CC) view where X-ray from top to bottom and mediolateral oblique (MLO) view where X-ray projects outward and downward at 45 degrees from the inner and upper part of the breast. Such dual-view mammogram is necessary and sufficient for radiologists to fully understand the 3D breast with 2D X-ray images, and thus can give accurate clinical decisions by following the standard Breast Imaging Reporting and Data System (BI-RADS) [Liberman and Menell, 2002]. However, mammogram inspecting is time-consuming and expertise-required, and usually suffers from intra- and inter-observer bias [Bae *et al.*, 2014]. Therefore, varied computer-aided diagnosis (CAD) systems have been emerging in recent years to provide fast and objective clinical decisions to assist large-scale screening.

Most traditional CADs [El-Naqa *et al.*, 2002; Arevalo *et al.*, 2016] for breast cancer screening rely on the analysis of individual lesions, and share a common pipeline mainly consisting of three consecutive steps (i.e., lesion detection, feature extraction, classification). They usually require a costly manual labeling of lesion masks either in training or test phase, heavily preventing the screening for a large population. In comparison, several advanced CADs have been proposed to classify a whole mammogram in a *holistic* fashion with only needs of the image-level supervisions indicating whether a mammogram contains malignant lesions or not. For exampel, Zhu *et al.* proposed the Deep MIL [Zhu *et al.*, 2017] to form the holistic mammogram classification problem into a multiple instance learning (MIL) task, and used three different MIL losses, i.e., max pooling loss, label assign loss, and sparsity loss, to fine-tune a pre-trained AlexNet. Similarly, Shu *et al.* [Shu *et al.*, 2020] designed region-based group-max pooling (RGP) and global-based group-max pool-

ing (GGP) to select more representative local features for the holistic mammogram classification.

Despite their success, these CADs can all be categorized as the single-view based approach, which treats CC and MLO views independently. However, the dual-view mammogram is naturally more suitable and useful than one-view information for reliable diagnosis. In a clinical practice, radiologists often resort to the MLO view to confirm the suspect lesions found in the CC view. Furthermore, a standard imaging routine typically provides a paired CC and MLO views for screening, bringing little extra workload for data acquisition to develop dual-view CADs. In view of these, many dual-view based CADs [AlGhamdi and Abdel-Mottaleb, 2021; Yan *et al.*, 2021b; Xian *et al.*, 2021; Cao *et al.*, 2021] have been emerged in the last decade.

Most existing dual-view based CADs make use of dual-view information to boost the performance of lesion detection, which is an intermediate task of breast cancer screening. For example, Yan *et al.* [Yan *et al.*, 2021a] utilized a shared YOLOv3 [Redmon and Farhadi, 2018] for proposing mass candidates and then paired each candidate across views and concatenated their features to directly classify whether they are matched or not by a metric network. Ma *et al.* [Ma *et al.*, 2021] proposed the Cross-View Relation Region-based CNN (CVR-RCNN) for robust mass detection by relating each candidate in one view to all candidates in another based on their feature similarities in order to better suppress false detections. For the task of classification, Bekker *et al.* [Bekker *et al.*, 2015] demonstrated a promising improvement for classifying lesions of clustered microcalcification (MC) by combining the results of two single-view logistic regressions on CC and MLO respectively. Carneiro *et al.* [Carneiro *et al.*, 2015; Carneiro *et al.*, 2017] combined images from both views associating with their corresponding lesion masks, i.e., mass and MCs, and explored how to fuse and where to fuse those dual-view information in a dual-path CNN.

Although the above dual-view CADs benefit from auxiliary information brought by additional views, they mostly learn features for different views independently, and produce the clinical decision by a simple combination of them (e.g., adding or concatenating). With no specific constraints, the underlying feature correlations (i.e., consistency and complementarity) across views are often ignored or failed to be captured, which leaves a great improving room. Furthermore, the holistic mammogram classification remains unstudied, and these CADs all require a prior of lesion masks, seriously hindering the application in large-scale screening.

In this paper, we for the first time aim to explicitly maximize the feature correlation across views for robust holistic mammogram classification requiring no lesion masks. The naive idea is to utilize a shared convolutional neural network (CNN) to extract single-view feature maps in parallel, and maximize the correlation loss proposed in [Yao *et al.*, 2017] to force the feature maps to be consensus across views. However, such correlation maximization has a *prerequisite* that the two input images are spatially aligned. For mammograms in our case, view changing and tissue superimposition make the dual-view images hardly meet the prerequisite, nullifying the benefits of the correlation maximization consequently.

To address this, there are two simple solutions. One is to spatially align pixels in the CC and MLO views, but practically infeasible. Another is to reduce spatial dimensions to loosen the requirement of alignment, but inevitably causes non-trivial information loss. Between these two solutions, we innovatively find a compromise that having each pixel enriched with information from its neighbors, which we argue is equivalent to local spatial dimension reduction but without much information loss. To this end, we empower the shared CNN by introducing both non-local and local attention mechanisms, and thus name it Dual-view Correlation Hybrid Attention Network (DCHA-Net)[1].

Concretely, the DCHA-Net has two shared branches for CC and MLO views respectively and each branch is a modified truncated ResNet101 [He *et al.*, 2016] with the last few bottlenecks replaced by the proposed hybrid attention module to reinvent features for the purpose of correlation maximization. The hybrid attention module consists of a local relation block [Hu *et al.*, 2019] and a non-local attention block [Wang *et al.*, 2018] to have each pixel in the resulting feature map contain information from its surroundings (local relation) as well as information of other pixels within its belonging strip-like region parallel to the chest wall (non-local attention). The motivation is based on a physical fact that two strip-like regions at the same distance from the chest wall are from the same tissue slice and thus matched and high-correlated, that is, *the CC and MLO views are roughly aligned along the direction perpendicular to the chest wall*. In view of this, the correlation loss is calculated within every matched strip-like regions in CC and MLO views, and optimized to make the two branches of DCHA-Net mutually assist each other.

In summary, our contributions are listed:

- We for the first time propose to learn dual-view features of mammograms by explicitly maximizing the correlations between those matched strip-like regions across views. With such constraint, the consistent and complementary dual-view features could be better captured even under no supervision of lesion labels, yielding a robust performance of holistic mammogram classification.
- We propose a DCHA-Net where the hybrid attention module enriches each pixel with its local contexts and global information of its belonging strip-like region, making the correlation maximization correct and effective even if the paired views are not aligned.
- We evaluate the proposed DCHA-Net on two well-known datasets, i.e., INbreast [Lee *et al.*, 2017] and CBIS-DDSM [Moreira *et al.*, 2012], and the experimental results verify our superior performance over other state-of-the-art methods of breast cancer classification.

2 Method

This section is organized as follows: we first describe the framework of DCHA-Net and how to maximize dual-view correlations in Sec. 2.1, then explain two naive solutions to meet the requirement of correlation maximization and lead

[1]https://github.com/BryantGary/IJCAI23-Dual-view-Correlation-Hybrid-Attention-Network.

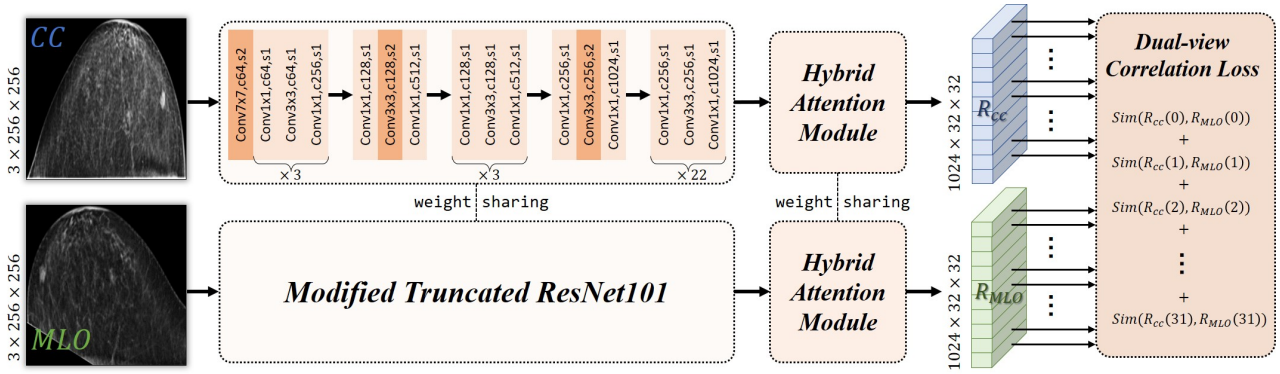


Figure 1: The framework of our proposed DCHA-Net which utilizes a shared modified truncated ResNet101 (without drawing the skip connections) for feature extraction and a shared hybrid attention module to reinvent feature maps for dual-view correlation maximization.$convK \times K, cN, sM$ means convolving by N kernels with the size of $K \times K$ and stride of M.

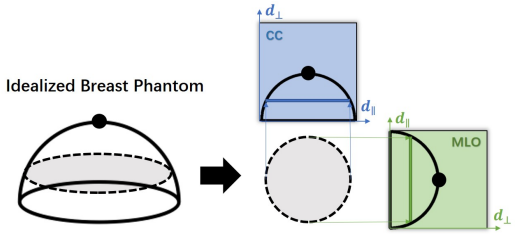


Figure 2: An idealized model explaining the relationship between 3D breast and 2D dual-view images. $\boldsymbol{d}_\parallel$ and $\boldsymbol{d}_\perp$ represent two directions parallel and perpendicular to the chest wall respectively.

out our solution in Sec. 2.2. At last, we detail the hybrid attention module in Sec. 2.3.

2.1 DCHA-Net and Correlation Maximization

Fig. 1 visualizes DCHA-Net, which contains two shared branches for CC and MLO view images, respectively. Given two images, i.e., I_{CC} and I_{MLO}, we first resize them to 256×256, remove backgrounds and pectoralis muscles, and then align the chest wall with the bottom edge of the image.

In each branch, we utilize a modified and truncated ResNet101 [He *et al.*, 2016] as a feature extractor, which is detailed in the light orange dashed box of Fig. 1. Compared to the vanilla ResNet101, we abandon the first max pooling layer and the last three bottlenecks (9 layers) to preserve more spatial information and to compromise computational costs for the usage of our proposed hybrid attention module. By three downsampling layers with the stride of 2, the feature extractor downscales the image by 8^2 times, yielding feature maps F_{CC} and F_{MLO} with the size of 32×32. The feature map is then reinvented by our proposed hybrid attention module to a new feature map, that is, $F_{CC} \rightarrow R_{CC}$ and $F_{MLO} \rightarrow R_{MLO}$, (see Sec. 2.3 for details).

A dual-view correlation loss inspired by [Yao *et al.*, 2017] is employed to explicitly maximize the feature correlations between the paired feature maps. Viewing the breast as a rigid semi-sphere, each slice in the 3D breast corresponds to two strip-like regions in both CC and MLO with equal distance to the chest wall as shown in Fig. 2, and the matched strip-like regions across views are thus highly-correlated with each other. Therefore, we propose to maximize correlations between every two row vectors with identical indexes, i.e., $R_{CC}(i)$ and $R_{MLO}(i)$, since their receptive fields just fit two matched strip-like regions. Concretely, the dual-view correlation loss is calculated as an average of cosine similarities between every matched row vectors in dual-view feature maps:

$$L_{corr} = -\frac{1}{32}\sum_{i=0}^{31} Sim\left(R_{CC}(i), R_{MLO}(i)\right) \tag{1}$$

where $R_{CC}(i)$ indicates the i-th row vector in R_{CC} and similar to $R_{MLO}(i)$. The cosine similarity $Sim\left(X, Y\right)$ is calculated as follows:

$$Sim\left(X, Y\right) = \frac{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\|X - \overline{X}\|\|Y - \overline{Y}\|} \tag{2}$$

where $\overline{X}$ is a scalar by averaging X and similar to $\overline{Y}$.

Note that the dual-view correlation loss in Eq. (1) is computed based on R_{CC} and R_{MLO} rather than the original feature maps F_{CC} and F_{MLO} because pixels are not aligned across views for the soft and non-rigid breast in the real world, not an idealized phantom shown in Fig. 2. Hence, forcibly calculating $Sim(F_{CC}, F_{MLO})$ could mess up the dual-view correlation loss and make it unable to give full effect. That is the very reason we introduce the hybrid attention module for feature reinvention. Before giving details of the hybrid attention module, we highlight our motivations in the next subsection.

2.2 Motivation of Hybrid Attention

The correlation loss is meant to enhance the feature learning for the multi-phase [Zhou *et al.*, 2019] or multi-modality [Yao *et al.*, 2017] data, while remains under-studied for the multi-view mammograms in our case. The main resistance is that the correlation loss asks the inputs should be spatially aligned beforehand, that is, $X(i, j)$ and $Y(i, j)$ in Eq. 2 should correspond to the same tissue for the same location (i, j). To this end, the very straightforward solution is to perform *registration* on dual-view images, which, however, is infeasible since

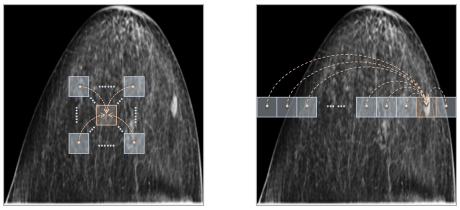


Figure 3: Diagrams of local (left) and non-local (right) attention mechanisms on a CC view image.

it is an ill-posed problem to completely disentangle those superimposed tissues from only two X-ray images according to the radon transform theory [Helgason and Helgason, 1980].

Another naive solution is *spatial dimension reduction*, performing global average pooling on F_{CC} and F_{MLO}, which makes correlation only rely on information along channels. However, this could result in dramatic information loss, since spatial dimension reduction is equivalent to eliminating feature differences between pixels at different locations.

Between these two simple solutions, we come up with a trade-off to alleviate the misalignment problem in dual-view mammograms. Instead of simply reducing the spatial dimensions, we introduce a local attention mechanism to make each pixel perceive its neighbors within a certain range, as shown in the left image of Fig. 3. Within the range, the misalignment has a chance to be corrected. That is, if it has a high possibility that a pixel $F_{MLO}(i,j)$ corresponds to its counterpart around $F_{CC}(i,j)$, we relate it to its belonging local patch $\{F_{CC}(i+\theta, j+\theta), -\sigma < \theta < \sigma, \sigma > 0\}$, where σ is the misalignment range, yielding a reinvented map $R_{CC}(i,j)$ (similar to $R_{MLO}(i,j)$). Hence, the reinvented features are more friendly to compute the dual-view correlation loss, since each pixel is already encoded with its neighbors, among which the corresponding pixel in another view can find its align one.

For pixels in each row across views, the misalignment range σ is hard to estimate, as implied in Fig. 2. To tackle this, we also introduce a non-local attention mechanism to have each pixel contain the entire information of its belonging row, as shown in the right image of Fig. 3. Combining both local and non-local attentions, the feature map extracted by the modified truncated ResNet101 is reinvented in a hybrid attention fashion. In the next subsection, we instantiate the hybrid attention as our proposed hybrid attention module and give details of its two key constitutions, i.e., local relation block and non-local attention block.

2.3 Hybrid Attention Module

A vanilla attention block [Vaswani *et al.*, 2017] typically contains three numerical units, i.e., a query of feature $q \in \mathbb{R}^{C\times 1}$ to encode, features $V \in \mathbb{R}^{C\times D}$ to relate, and keys of features $K \in \mathbb{R}^{C\times D}$ to compute attentions with the query. The relations are obtained by performing the dot products of the query with all keys, dividing each by $\sqrt{C}$, and applying a softmax function successively. Hence, we have a new feature $f' \in \mathbb{R}^{C\times 1}$ encoded with all information from V based on the attention block:

$$f'^T = softmax(\frac{q^T K}{\sqrt{C}})V^T \tag{3}$$

Next, we describe the two instantiations of Eq. (3), i.e., local relation block and non-local attention block, which are designed to accomplish the manipulations visualized in Fig. 3.

Local Relation Block

As shown in the left part of Fig. 4, we first compute a key pool and a query pool from the original feature map $F \in \mathbb{R}^{C\times H\times W}$ (omitting the subscripts CC and MLO) by two 1×1 convolution layers. C indicates the dimensions of the channel and $H \times W$ indicates the feature map size.

From the query pool denoted as the orange cuboid in Fig. 4, we extract feature vectors as queries at every location by a 1×1 sliding window (also can directly reshape), forming HW queries $\{q_i \in \mathbb{R}^{C\times 1}, i = 0, ..., HW-1\}$. Similarly, we can have keys $\{K_i \in \mathbb{R}^{C\times k^2}, i = 0, ..., HW-1\}$ by performing a $k\times k$ sliding window with zero padding from the key pool denoted as the green cuboid in Fig. 4 and by flattening. The parameter k controls the range of the possible misalignment, and we empirically set it to 3 corresponding to a 24×24 receptive field in the original image I.

Each relation map is obtained by $softmax(q_i^T K_i/\sqrt{C})$, and thus those related features $\{f'_i \in \mathbb{R}^{C\times 1}, i = 0, ..., HW-1\}$ can be calculated as:

$$f_i'^T = softmax(\frac{q_i^T K_i}{\sqrt{C}})V_i^T \tag{4}$$

where features $\{V_i \in \mathbb{R}^{C\times k^2}, i = 0, ..., HW-1\}$ are extracted also by the sliding window and flattening on the original feature map F. Those related features with the total number of HW are at last packed together to $F' \in \mathbb{R}^{C\times H\times W}$ by the reverse sliding window process. A skip connection is employed to have $F' = F' + F$.

By doing so, each entry of F' carries information from its corresponding pixel in F, and neighboring ones defined by V. Furthermore, the relation map defined based on q_i and K_i is also learned to make each entry in F' pay more attention to its surroundings more likely to compensate misalignment to another view, and vice versa.

Non-local Attention Block

As shown in the right part of Fig. 4, we also compute a key pool and a query pool from F' by another two 1×1 convolution layers.

Unlike the local relation block, we generate keys $\{K_i \in \mathbb{R}^{C\times W}\}$, queries $\{Q_i \in \mathbb{R}^{C\times W}\}$, and features $\{V_i \in \mathbb{R}^{C\times W}\}$ by a $1\times W$ sliding window, where $i = 0, ..., H-1$, since we intend to relate each pixel to its belonging strip-like region (i.e., each row). Note that, Q_i is also equivalent to packing together all queries q_i belonging to the same row. Hence, the related features $\{F''_i \in \mathbb{R}^{C\times W}, i = 0, ..., H-1\}$ can be calculated as:

$$F_i''^T = softmax(\frac{Q_i^T K_i}{\sqrt{C}})V_i^T \tag{5}$$

The overall attention-derived feature map R is obtained by a reverse sliding window on $\{F''_i\}$. Similarly, a skip connection is employed to have $R = R + F'$.

Empowered by the hybrid attention module, R has each entry related to both its strip-like region and neighbors, and

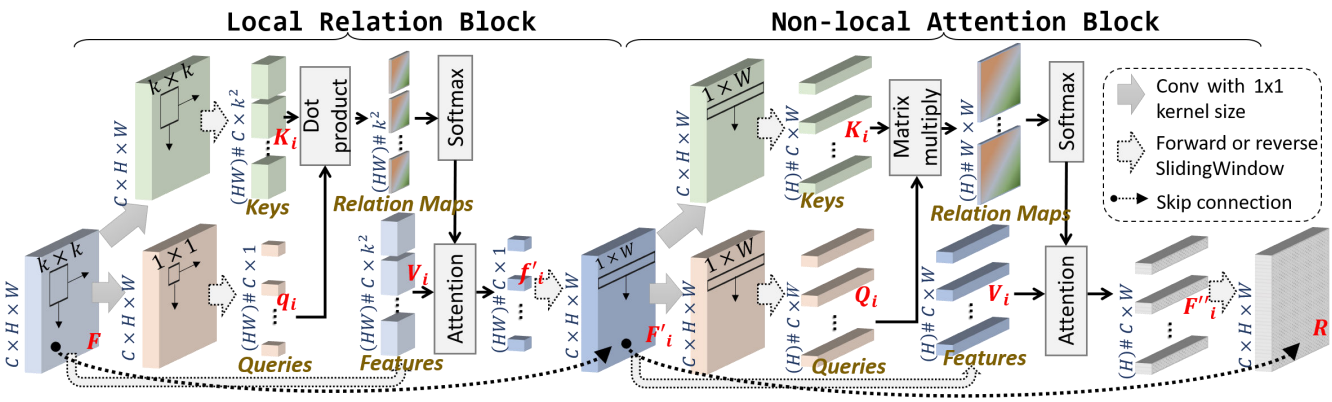


Figure 4: Detailed architectures of two key blocks of our proposed DCHA-Net, i.e., local relation block and non-local attention block. $(HW)\#\ C \times k^2$ means numerical units with the total number of HW and the size of $C \times k^2$, and similar to others.

thus makes the dual-view correlation loss in Eq. (1) play its full effect, better mining and preserving those underlying feature correlations across views without a need of registration.

2.4 Loss for Training

We contribute the proposed DCHA-Net to solve the holistic mammogram classification task. To this end, we add two classification heads without weight sharing on the top of the extracted and reinvented feature maps, i.e., R_{CC} and R_{MLO} respectively, whose size is $1024 \times 32 \times 32$ as shown in Fig. 1. Each head consisting of two layers first performs global average pooling to get a 1024-d representation, and utilizes a full-connected layer and sigmoid function to predict a single unit p indicating the possibility of the input image containing the malignant breast tumor lesion or not. The classification losses for the two views are calculated as follows:

$$L_{clss}^{CC} = -\left(ylog(p_{CC}) + (1-y)log(1-p_{CC})\right) \quad (6)$$

$$L_{clss}^{MLO} = -\left(ylog(p_{MLO}) + (1-y)log(1-p_{MLO})\right) \quad (7)$$

where p_{CC} and p_{MLO} are predictions for CC and MLO view images respectively. Totally, the final loss to train DCHA-Net is thus calculated as:

$$L = L_{corr} + L_{clss}^{CC} + L_{clss}^{MLO} \quad (8)$$

3 Experiments

3.1 Datasets and Experimental Setup

INbreast: The INbreast dataset collects in total 410 full-field digital mammographic images, from which 90 cases, i.e., patients, with both breasts and 25 cases with only one side of breast are included. Multiple different types of annotations are provided, including the BIRADS classification scores, mass/calcification masks for segmentation, and other annotations such as pectoralis muscles and distortions.

CBIS-DDSM: The CBIS-DDSM dataset has in total 3071 scanned film mammography images (including 891 mass cases and 753 calcification cases). The CBIS-DDSM is a selected version of DDSM, with higher image and label quality, and more friendly access. It also contains precise annotations including ROI segmentation masks, bounding boxes and the BIRADS scores.

We label images as two classes: the BIRADS scores belonging to $\{1, 2, 3\}$ as normal or benign, and $\{4, 5, 6\}$ as malignant. For INbreast, we randomly split 80% cases for training and the remaining 20% cases for test. For CBIS-DDSM, we follow its default division setting with 85% training cases and 15% test cases. Note that, no images from the same patient are cross-used in training and test sets for the two datasets. During the training, we exclude cases with only single view, and augment the original data with random rotation and flipping. For the evaluation, we utilize two metrics, i.e., the Accuracy and the Area Under the receiver operating characteristic Curve (ROC), i.e., the AUC value.

3.2 Implementation Details

For data pre-processing, we first use OpenCV edge detection to remove background. We use the provided GT masks to remove pectoralis muscles for INbreast. We manually find a line fitting the chest wall, and then remove regions on the non-breast side for CBIS-DDSM. We use lines to fit chest wall and align the two lines across views.

The proposed DCHA-Net is implemented with the Pytorch library and trained on one NVIDIA GeForce RTX 3090 GPU with 24 GB memory. We use a pretrained ResNet-101 for weight initialization, and an Adam optimizer. The learning rate starts at $5e$-5 and gradually decays by 0.9. Our method uses two classification heads and thus gives two predictions for a dual-view image of a breast, and we take the average output unit as the final predicted probability of being malignant and binarize it using a threshold 0.5.

3.3 Comparison with the State-of-the-arts

We first compare our method with eight previous state-of-the-arts on INbreast, including Domingues *et al.* [Domingues *et al.*, 2012], deep MIL [Zhu *et al.*, 2017], Shams *et al.* [Shams *et al.*, 2018], RGP and GGP [Shu *et al.*, 2020], Carneiro *et al.* [Carneiro *et al.*, 2017] and MCRLA [Li *et al.*, 2021]. As shown in Table 1, most approaches are single view-based methods, which can hardly achieve satisfactory performance without extracting and utilizing dual-view informa-

Method	Views	Data Division	Accuracy	AUC
Dataset: INbreast				
Domingues *et al.* [Domingues *et al.*, 2012]	Single	Image	0.890	-
Pretrained CNN+RF [Dhungel *et al.*, 2016]	Single	Image	0.910±0.02	0.760±0.23
Deep MIL [Zhu *et al.*, 2017]	Single	Image	0.900±0.02	0.890±0.04
Shams *et al.* [Shams *et al.*, 2018]	Single	Image	0.935±0.03	0.925±0.02
RGP [Shu *et al.*, 2020]	Single	Image	0.919±0.03	0.934±0.03
GGP [Shu *et al.*, 2020]	Single	Image	0.922±0.02	0.924±0.03
Carneiro *et al.* [Carneiro *et al.*, 2017]	Dual	Patient	-	0.860±0.09
MCRLA [Li *et al.*, 2021]	Dual	Patient	0.912	0.942
DCHA-Net	Dual	Patient	**0.955±0.01**	**0.950±0.02**
Dataset: CBIS-DDSM				
Deep MIL [Zhu *et al.*, 2017]	Single	Patient	0.742±0.03	0.791±0.02
RGP [Shu *et al.*, 2020]	Single	Patient	0.762±0.02	0.838±0.01
GGP [Shu *et al.*, 2020]	Single	Patient	0.767±0.02	0.823±0.02
MCRLA [Li *et al.*, 2021]	Dual	Patient	0.766	0.824
Petrini *et al.** [Petrini *et al.*, 2022]	Dual	Patient	-	0.842±0.03
DCHA-Net	Dual	Patient	**0.781±0.01**	**0.846±0.01**

Table 1: Quantitative comparison of different methods on both the INbreast and the CBIS-DDSM dataset. Our final results of DCHA-Net are obtained by using data-augmentation techniques during training. Results of other groups except ours are directly inherited from papers of Shams *et al.*, RGP/GGP, Carneiro *et al.*, MCRLA and Petrini *et al.*. '*' indicates results without test-time augmentation for a fair comparison.

tion. Moreover, deep MIL, Shams *et al.* and RGP/GGP all divide data by images, where images in the same case may be split into the training and testing set at the same time. In contrast, DCHA-Net effectively mines dual-view features on patient-division data, which surpasses all state-of-the-art approaches by a great margin, achieving the best results in terms of the average Accuracy (0.955) and the AUC (0.950).

We also evaluate on the CBIS-DDSM dataset. As shown in the bottom of Table 1, we compare our DCHA-Net with Deep MIL, RGP/GGP, MCRLA and Petrini *et al.* [Petrini *et al.*, 2022]. Note that all approaches followed the default data division in the CBIS-DDSM dataset and trained on patient-division data. In comparison, our approach remarkably achieves the best average accuracy and the AUC value, remaining consistency results as those on INbreast.

Comparing results between the two datasets, it is worth noting that the performance on the INbreast dataset is greater than that on the CBIS-DDSM dataset. We believe this is possibly caused by different image quality. For instance, the INbreast images were collected using more advanced mammography screening techniques, which can help extract more useful features during training.

3.4 Ablation Analysis

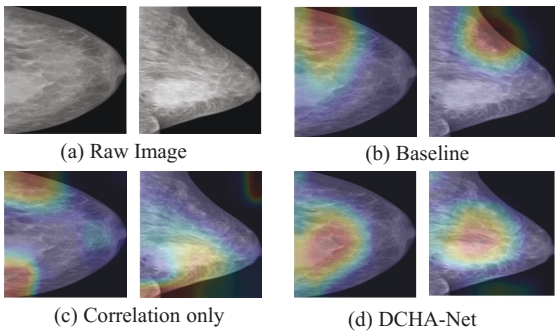
Ablation Analysis of Key Components

We conduct an ablation study on the INbreast dataset to analysis impacts of key components in the DCHA-Net. Here, we disable the data augmentation techniques used in comparison with the state-of-the-arts. Table 2 shows the comparison results of six variants, including: 1) "Baseline" which directly trains on clear ResNet backbones (see the 1^{st} row); 2) "Corr. only" that only utilizes correlation constraints during training (the 2^{nd} row); 3) "Corr. plus local relation" that uses only local attention with correlation constraints (the 3^{rd} row); 4) "Corr. plus non-local atten." that uses only non-local attention with correlation constraints (the 4^{th} row); 5) "Hybrid-atten. only" that only adds the hybrid-attention module (the 5^{th} row) and 6) "DCHA-Net" that utilizes both the hybrid-attention module and correlation constraints (the 6^{th} row).

Four observations can be made from the results. First, the baseline achieves the worst performance, indicating the significance of both the correlation constraints and the hybrid-attention module. More specifically, the correlation constraints and the hybrid-attention module can respectively result in an increment of 1.136% and 1.515% in the average accuracy, and an increment of 0.007 and 0.043 in the AUC value. Second, compared to Baseline, DCHA-Net greatly improves the accuracy by 4.924% and the AUC by 0.067. Third, when only using local (non-local) attention with correlation maximization, ACC is 90.152% (89.773%) and AUC is 0.909 (0.901), inferior to those achieved by combining both attentions (i.e., hybrid). Fourth, we can observe that the correlation constraints can result in higher improvements after using the hybrid-attention module. This indicates that the hybrid-attention module can effectively tackle the dual-view spatial misalignment problem, and help capture correct correlation information maximally.

In addition, we also perform student t-test, and report p-values for both metrics. Comparing DCHA-Net with Baseline (Correlation-only), the p-values for Accuracy and AUC are $7.78e$-4 and $2.81e$-2 ($1.11e$-2 and $1.05e$-3), respectively. Comparing Corr. only with Baseline, the corresponding p-values are $2.51e$-1 and $7.66e$-1. This indicates the effectiveness of the hybrid-attention module on misalignment.

To further demonstrate the improvements of mining dual-view information, we used Grad-CAM [Selvaraju *et al.*, 2017] to visualize the most suspicious malignant areas (e.g., mass) predicted by different groups of methods. Grad-CAM uses the gradient as weight to highlight attentive areas, which are those more contribute to the final classification prediction.



(a) Raw Image (b) Baseline (c) Correlation only (d) DCHA-Net

Figure 5: Visualization of the most suspicious malignant areas predicted by different methods. Each group is sampled from a same case with paired CC and MLO images.

Local atten.	Non-local atten.	Dual-view corr.	Acc. (%)	AUC
✗	✗	✗	87.879	0.870
✗	✗	✓	89.015	0.877
✓	✗	✓	90.152	0.909
✗	✓	✓	89.773	0.901
✓	✓	✗	89.394	0.913
✓	✓	✓	**92.803**	**0.937**

Table 2: Ablation analysis of the proposed hybrid-attetion module and employed correlation maximization.

Visual sign of lesions is mostly related to classification label. As shown in Fig. 5(b), without correlation loss, the gradient cannot flow across views, making the two view features unable to "cross-check" and easily distracted by some lesion-irrelevant regions. Therefore, the baseline model leans to focus on lesion-irrelative and cross-view mismatched areas. With correlation only, the two view features cannot "cross-check" at the truly matched places, bringing incorrect and confusing information from another view due to the spatial misalignment problem (see Fig. 5(c)). As shown in Fig. 5(d), with correlation plus hybrid-attention, the misalignment is alleviated in the feature space, and the matched lesion-relevant regions can be successfully highlighted by Grad-CAM.

Effectiveness of Hybrid-attention Module

The hybrid-attention module consists of two basic blocks, i.e., the local attention block and the non-local attention block. We conduct an ablation study of their effectiveness on INbreast using two different settings, i.e., "mixed views" and "single view", and the data augmentation techniques are also disabled. Under "mixed views", we mix images from the two views together and simply train our model by using a single shared classification head without constraining the dual-view correlation. Under "single view", we split the data into two parts, and each contains images from a single view. We independently train a model for each part, and thus the correlation constraint is naturally disabled.

We report the performance on each view for the two settings, and the results are shown in Table 3. As can be seen from the results, both the local-attention block and the non-local attention block can contribute to large improvements. For instance, under "mixed views", the local-attention block and the non-local attention block respectively result in an increment of 0.947% and 1.137% in the average accuracy. In conjunction of these two components, the hybrid-attention module achieves the best performance, e.g., improving the average accuracy by 2.273% and 1.516% respectively under these two settings. This also implies that the two components can work synthetically to result in a more robust performance.

Training view	Local Relation Block	Non-Local Attention Block	Accuracy (%) on CC view	Accuracy (%) on MLO view
Mixed views	✗	✗	88.636	87.879
	✓	✗	89.773	88.636
	✗	✓	89.394	89.394
	✓	✓	**90.152**	**90.909**
Single view	✗	✗	88.636	87.121
	✓	✗	89.394	87.879
	✗	✓	89.394	87.879
	✓	✓	**90.152**	**88.636**

Table 3: Ablation study of the effectiveness of the two different attention blocks in hybrid attention module under two different settings of images for training. The correlation maximization is disabled for verifying the hybrid attention purely.

4 Conclusion

In this paper, we propose a novel end-to-end DCHA-Net which contains two key components for robust holistic mammographic classification. First, the dual-view correlation loss aims at maximizing paired feature similarity across two views, which effectively helps capture consistent and complementary information for better mammographic classification accuracy. In addition, the hybrid-attention module reinvents information from local and strip-like non-local regions into every pixel, alleviating negative influences brought by the spatial misalignment problem and guaranteeing the extracted dual-view correlated features correct. Extensive experimental results on both the INbreast and CBIS-DDSM datasets demonstrate that our proposed DCHA-Net can significantly improve the breast cancer diagnosis performance and outperform previous state-of-the-art methods.

Contribution Statement

Zhiwei Wang and Junlin Xian are the co-first authors contributing equally to this work. Qiang Li and Xin Yang are corresponding authors.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Grant No. 62202189), Fundamental Research Funds for the Central Universities (2021XXJS033), research grants from Wuhan United Imaging Healthcare Surgical Technology Co., Ltd.

References

[AlGhamdi and Abdel-Mottaleb, 2021] Manal AlGhamdi and Mohamed Abdel-Mottaleb. Dv-dcnn: Dual-view deep convolutional neural network for matching detected masses in mammograms. *Computer methods and programs in biomedicine*, 207:106152, 2021.

[Arevalo *et al.*, 2016] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016.

[Bae *et al.*, 2014] Min Sun Bae, Woo Kyung Moon, Jung Min Chang, Hye Ryoung Koo, Won Hwa Kim, Nariya Cho, Ann Yi, Bo La Yun, Su Hyun Lee, Mi Young Kim, et al. Breast cancer detected with screening us: reasons for nondetection at mammography. *Radiology*, 270(2):369–377, 2014.

[Bekker *et al.*, 2015] Alan Joseph Bekker, Moran Shalhon, Hayit Greenspan, and Jacob Goldberger. Multi-view probabilistic classification of breast microcalcifications. *IEEE Transactions on medical imaging*, 35(2):645–653, 2015.

[Cao *et al.*, 2021] Zhenjie Cao, Zhicheng Yang, Yuxing Tang, Yanbo Zhang, Mei Han, Jing Xiao, Jie Ma, and Peng Chang. Supervised contrastive pre-training formammographic triage screening models. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pages 129–139. Springer, 2021.

[Carneiro *et al.*, 2015] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

[Carneiro *et al.*, 2017] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE transactions on medical imaging*, 36(11):2355–2365, 2017.

[Dhungel *et al.*, 2016] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. The automated learning of deep features for breast mass classification from mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 106–114. Springer, 2016.

[Domingues *et al.*, 2012] I Domingues, E Sales, J Cardoso, and W Pereira. Inbreast-database masses characterization. *XXIII CBEB*, 2012.

[El-Naqa *et al.*, 2002] Issam El-Naqa, Yongyi Yang, Miles N Wernick, Nikolas P Galatsanos, and Robert M Nishikawa. A support vector machine approach for detection of microcalcifications. *IEEE transactions on medical imaging*, 21(12):1552–1563, 2002.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Helgason and Helgason, 1980] Sigurdur Helgason and S Helgason. *The radon transform*, volume 2. Springer, 1980.

[Hosseini *et al.*, 2016] Hedayatollah Hosseini, Milan MS Obradović, Martin Hoffmann, Kathryn L Harper, Maria Soledad Sosa, Melanie Werner-Klein, Lahiri Kanth Nanduri, Christian Werno, Carolin Ehrl, Matthias Maneck, et al. Early dissemination seeds metastasis in breast cancer. *Nature*, 540(7634):552–558, 2016.

[Hu *et al.*, 2019] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.

[Lee *et al.*, 2017] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017.

[Li *et al.*, 2021] Dong Li, Lituan Wang, Ting Hu, Lei Zhang, and Qing Lv. Deep multiinstance mammogram classification with region label assignment strategy and metric-based optimization. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4):1717–1728, 2021.

[Liberman and Menell, 2002] Laura Liberman and Jennifer H Menell. Breast imaging reporting and data system (bi-rads). *Radiologic Clinics*, 40(3):409–430, 2002.

[Lu *et al.*, 2018] Chia-Feng Lu, Fei-Ting Hsu, Kevin Li-Chun Hsieh, Yu-Chieh Jill Kao, Sho-Jen Cheng, Justin Bo-Kai Hsu, Ping-Huei Tsai, Ray-Jade Chen, Chao-Ching Huang, Yun Yen, et al. Machine learning–based radiomics for molecular subtyping of gliomas. *Clinical Cancer Research*, 24(18):4429–4436, 2018.

[Ma *et al.*, 2021] Jiechao Ma, Xiang Li, Hongwei Li, Ruixuan Wang, Bjoern Menze, and Wei-Shi Zheng. Cross-view relation networks for mammogram mass detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8632–8638. IEEE, 2021.

[Moreira *et al.*, 2012] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[Moss *et al.*, 2012] SM Moss, Lennarth Nyström, Hakan Jonsson, E Paci, E Lynge, S Njor, and M Broeders. The impact of mammographic screening on breast cancer mortality in europe: a review of trend studies. *Journal of medical screening*, 19(1_suppl):26–32, 2012.

[Petrini *et al.*, 2022] Daniel GP Petrini, Carlos Shimizu, Rosimeire A Roela, Gabriel Vansuita Valente, Maria Aparecida Azevedo Koike Folgueira, and Hae Yong Kim. Breast cancer diagnosis in two-view mammography using

end-to-end trained efficientnet-based convolutional network. *Ieee Access*, 10:77723–77731, 2022.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Shams *et al.*, 2018] Shayan Shams, Richard Platania, Jian Zhang, Joohyun Kim, Kisung Lee, and Seung-Jong Park. Deep generative breast cancer screening and diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 859–867. Springer, 2018.

[Shu *et al.*, 2020] Xin Shu, Lei Zhang, Zizhou Wang, Qing Lv, and Zhang Yi. Deep neural networks with region-based pooling structures for mammographic image classification. *IEEE Transactions on Medical Imaging*, 39(6):2246–2255, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[Xian *et al.*, 2021] Junlin Xian, Zhiwei Wang, Kwang-Ting Cheng, and Xin Yang. Towards robust dual-view transformation via densifying sparse supervision for mammography lesion matching. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 355–365. Springer, 2021.

[Yan *et al.*, 2021a] Yutong Yan, Pierre-Henri Conze, Mathieu Lamard, Gwenolé Quellec, Béatrice Cochener, and Gouenou Coatrieux. Towards improved breast mass detection using dual-view mammogram matching. *Medical image analysis*, 71:102083, 2021.

[Yan *et al.*, 2021b] Yutong Yan, Pierre-Henri Conze, Mathieu Lamard, Heng Zhang, Gwenolé Quellec, Béatrice Cochener, and Gouenou Coatrieux. Deep active learning for dual-view mammogram analysis. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 180–189. Springer, 2021.

[Yao *et al.*, 2017] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 406–414. Springer, 2017.

[Zhou *et al.*, 2019] Yuyin Zhou, Yingwei Li, Zhishuai Zhang, Yan Wang, Angtian Wang, Elliot K Fishman, Alan L Yuille, and Seyoun Park. Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 155–163. Springer, 2019.

[Zhu *et al.*, 2017] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2017.