Towards Robust Natural-Looking Mammography Lesion Synthesis on Ipsilateral Dual-Views Breast Cancer Analysis

Thanh-Huy Nguyen[1,3], Quang Hien Kha[2,3], Thai Ngoc Toan Truong[4], Ba Thinh Lam[3], Ba Hung Ngo[5], Quang Vinh Dinh[6], and Nguyen Quoc Khanh Le[2, 7, 8]

[1]Department of Biomedical Engineering, National Cheng Kung University, Taiwan
[2]International Ph.D. Program in Medicine, College of Medicine, Taipei Medical University, Taiwan
[3]Saigon Precision Medicine Research Center (SAIGONMEC), Vietnam
[4]Ho Chi Minh City International University, Vietnam
[5]Graduate School of Data Science, Chonnam National University, Korea
[6]Vietnamese-German University, Vietnam
[7]Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taiwan
[8]Professional Master Program in Artificial Intelligence in Medicine, Taiwan

Abstract

In recent years, many mammographic image analysis methods have been introduced for improving cancer classification tasks. Two major issues of mammogram classification tasks are leveraging multi-view mammographic information and class-imbalance handling. In the first problem, many multi-view methods have been released for concatenating features of two or more views for the training and inference stage. Having said that, most multi-view existing methods are not explainable in the meaning of feature fusion, and treat many views equally for diagnosing. Our work aims to propose a simple but novel method for enhancing examined view (main view) by leveraging low-level feature information from the auxiliary view (ipsilateral view) before learning the high-level feature that contains the cancerous features. For the second issue, we also propose a simple but novel malignant mammogram synthesis framework for upsampling minor class samples. Our easy-to-implement and no-training framework has eliminated the current limitation of the CutMix algorithm which are unreliable synthesized images with random pasted patches, hard-contour problems, and domain shift problems. Our results on VinDr-Mammo and CMMD datasets show the effectiveness of our two new frameworks for both multi-view training and synthesizing mammographic images, outperforming the previous conventional methods in our experimental settings.

[*]Thanh-Huy Nguyen and Quang Hien Kha have equal contributions.

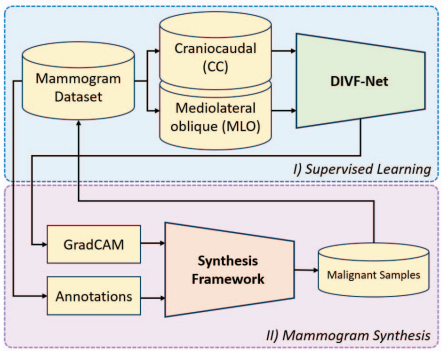

Figure 1. Our proposed pipeline for training and synthesizing mammographic images. Two stages are the supervised training on ipsilateral views mammograms. and synthesis framework that takes the saliency map and region malignant annotations.

1. Introduction

Breast cancer has one of the highest rates of mortality and incidence among women worldwide, making it one of the most common cancers to cause death. Cancer detection, in particular at the early stage, must be crucial in screening mammogram exams. Both the craniocaudal (CC) view and the mediolateral oblique (MLO) view, which are top-down and side views of the breast, respectively, can be used to classify each patient's breasts. Radiologists frequently examine both views of the same breast (ipsilateral views) and the same view of both breasts (bilateral views) to make a sound, intuitive decision.

Based on that, prior works nowadays can be classified into various groups: ipsilateral-only based, bilateral-only based, and ipsilateral-bilateral combination. Recent papers expose bilateral-only based, Liu et al. [16] enhanced mammogram mass detection using a contrasted bilateral network (CBN). Furthermore, Zhao et al. [32] used a well-known attention module between adaptive spatial and channel that yields the categorization. In contrast, those strategies still struggle with the conflict between the two breast sides that cause noise in the model because one patient might have the disease on one breast while another does not. Another group is the ipsilateral-bilateral combination approach used three views or four views as the inputs which create a full overview of breasts. Liu et al. [14, 15] achieved this by proposing a remarkable Anatomy-aware Graph Convolutional Network (AGN) that relies on the mass shape and region to construct the graphical correspondence among different mammographic views. Although the performance of these models is noteworthy, they require massive computational, thus might be hardly embedded in hospital facilities. Continuously, Nguyen et al. [21] proposed four views input, independently, each view will be learned to extract features and then fed into Light-GBM [10] classifier for prediction. Afterward, the result operates the max function between ipsilateral view sides, which can be inaccurate and lead to a poor learning process.

Furthermore, mammogram synthesis and augmentation techniques are also one of the most promising approaches for handling class imbalance. MixUp [31] reduces the proportion of informative pixels of two images to produce a new image that shows impressive results in many medical image applications. Similar to MixUp, the CutMix [30] algorithm generates a simple augmentation methodology that replaces the patch of two images together. However, both of these methods might cause a conflict in the label because the random of choosing a patch in the mammographic image can create two different classes in the same image. About MixUp, the algorithm itself uses the image-level mixing between two images without semantic label preserving for mammographic cancer classification. Besides, the region generated from CutMix is random, that might or might not contain the cancerous information when conducting copy-paste. CutMix also might create new untrustworthy samples due to the solid rectangle's boundary of pasted patches and the difference in style space.

To take full advantage, we propose a Dual Ipsilateral Views Fusion Network (DIVF-Net) for mammographic image classification. This network can be separated into three parts: Low-Level Feature Blocks, Features Fusion Blocks, and High-Level Feature Blocks. Our network can leverage low-level information such as the shape, contour, and density of the breasts. The DIVF-Net combines two low-level features for extracting the relevant information before using it for enhancing the main view feature. The high-level information part of DIVF-Net aims to focus on the lesions that highly contain semantic information for cancer classification. Additionally, a Malignant Lesions Synthesis Framework also is proposed in this paper which overcomes the current limitations of CutMix and MixUp algorithms. It includes three stages: Region Selection, Domain Adaptation, and Soft Contour Transformation. The framework carefully picks the radiologist-annotated region for replacing the benign-information-contained region. The rest of the framework aims to close the gap of different between source and target patches before replacing it with a gradient-contour MixUp algorithm.

In summary, the main contributions of our work are as follows:

- A novel multi-view network DIVF-Net with two types of fusion operations that leverages information on both CC and MLO views for accurate cancer classification.
- A new robust mammogram synthesis framework that replaces the benign to malignancy region with an informative region. The created patches are also being smoothed and Fourier-adapted before replacing the indicated regions.
- Experimental results and ablation studies based on a combination of these two show the robustness and generalizability on multiple fusion settings and datasets.

2. Related Work

Multi-view Network: Compared with 2D views, 3D objects have much more knowledge to guide the model, which is described in visual understanding [6, 24] and stereo vision [3, 4, 23]. In visual understanding, they set several cameras around a target object to model region-to-region and views from various angles. Each view is embedded in a shared weight Convolutional Neural Network (CNN). In stereo vision, two cameras are placed closely. This approach is mainly used in self-driving cars, which manipulate the depth estimation via disparity map fusion. The depth estimation helps the system knows the closeness to itself the straight object to immediately avoid the car collapse and keep a safe distance. Multi-view-based approaches [21, 26] collect features from various 2D views to represent the 3D object. First, they fed each view into a feature extractor to learn the appropriate embedding feature. Then, they proposed their work to significantly fuse all of them for 3D representation.

Inspired by that, mammographic screening also has a differentiated imaging process that is efficient to represent 3D objects. *Wu et al* [28] proposed the four views mammogram network to predict the malignant or not malignant classification. They aggregate between the bilateral views at the

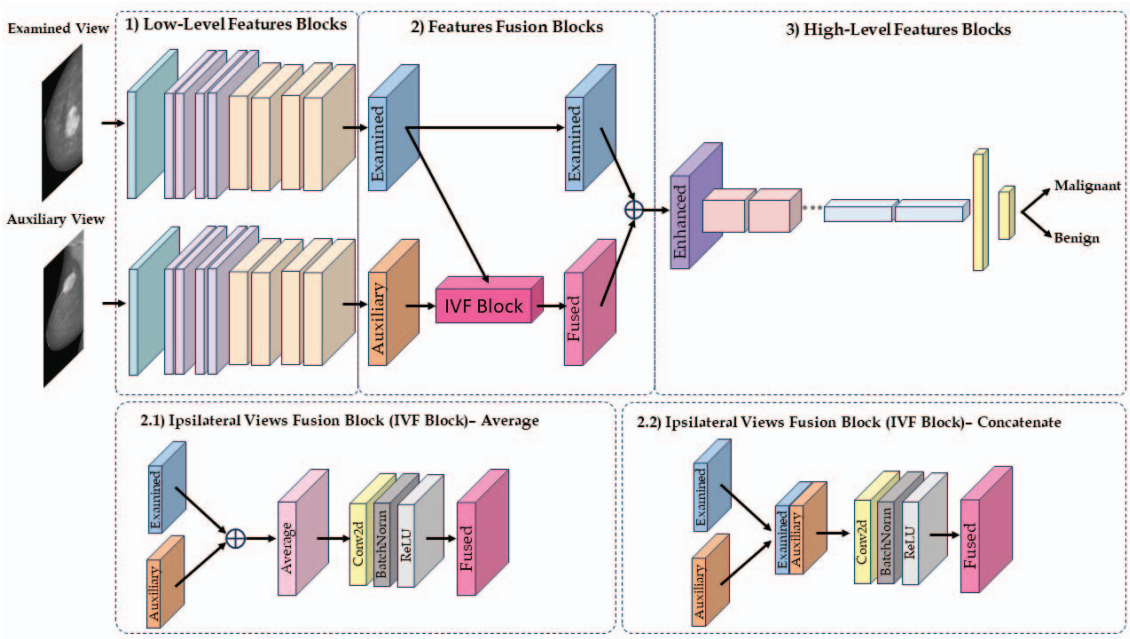

Figure 2. Dual Ipsilateral Views Fusion Network (DIVF-Net) for Mammographic Cancer Diagnosis. This framework consists of three stages: Low-Level Features Blocks (left-top), Features Fusion Blocks (middle-top), and High-Level Features Blocks(right-top). In Ipsilateral Views Fusion (IVF) Block, there are two operations used to fuse both examined features and auxiliary features: average and concatenate. After going through IVF Block, the fused features combine with examined features to improve the performance.

first stage and then the softmax layer. Finally, they presented four strategies with a combination of several layers. *Khan et al* [19] enhanced the way mammogram image preprocessing and decreased the computational complexity in the backbone. They extract directly a mass via augmented ROIs and modify a small VGGNet-like architecture used for the feature extraction stage. In general, ipsilateral views consist of CC and MLO views of the same breast side. This advantage in extracting rich information for 3D medical image analysis. Thus, fusing the ipsilateral views increases the global features in fusion operation beside the local features from individual views.

Medical Image Synthesis/Augmentation: Augmentation is one of the most fundamental procedures for synthesizing training data for further generalizability. Existing works on data augmentation [13, 30, 31] synthesize two images into soft images. Thus, the generated new training images direct the model to concentrate more on shape than texture, which improves classification and object identification performances. CutOut [5] revivals the object occlusion, which is a common issue in many computer vision tasks. It randomly chooses one defined size patch to remove. While CutMix [30] replaces the binary mask with another image and mixes the label via the combination ratio. Mixup [31] sampling from the mixup vicinal distribution produces virtual feature-target vectors.

In recent years, Generative Adversarial Networks (GAN) [7] become a well-known deep-learning-based medical image synthesis framework. For the synthesis of mammograms, Dimitrios Korkinof et al [12]. employ a progressive GAN (PGGAN), achieving high resolutions and positive outcomes when comparing the low-level pixel distributions of real and artificial images. Rui Man et al.'s research [18] focuses on creating synthetic samples, but in this instance, they create patches of a histopathological image. This AnoGAN (Anomaly Detection GAN) has many benefits for teaching classification systems. Xiangyuan Ma et al.'s [17] research focuses on creating samples of mammogram lesion segmentation masks. This enables overcoming image labeling, one of the most difficult tasks involved in dataset construction. Having said that, the biggest concern of GAN-based approaches is the realism and trustworthiness of synthesized samples. It may not be practical in real-world applications when using synthesized mammograms for training and testing.

3. Methodology

3.1. Dual Ipsilateral Views Fusion Network

This work aims to exploit the dual-view mammograms of the same breast using a new proposed network, DIVF-Net. Our network takes two ipsilateral views (CC and MLO) of a single breast to assess the cancerous. For each patient, the model takes one view of the breast as an examined view, and the other is an auxiliary view to support. As shown in Fig. 2, both examined view and auxiliary view

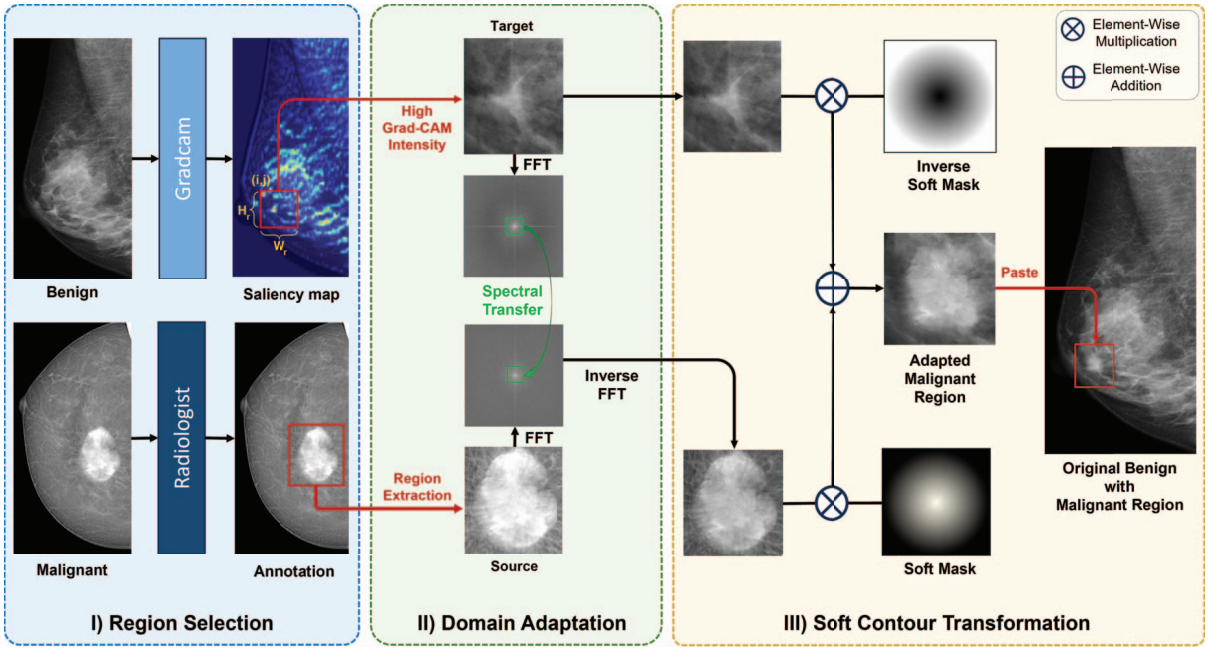

Figure 3. Proposed Soft-Adapted Malignancy Synthesis Framework consists of three phases: 1) Region Selection, extracting the important region (left); 2) Domain Adaptation, adapting target style to the source image (middle); 3) Soft Contour Transformation, smoothing the target region with an inverse soft mask and the transformed source region with a soft mask (right).

are fed into Low-Level Features Blocks (the first half of the popular backbone like ResNet [8]). Then, the output features of these views are combined by IVF Block. The IVF Block includes 4 components: Aggregation mechanism, 2D convolutional layer, batch normalization layer [9], and ReLU activation function [1]. In the aggregation part, there are two ways to combine two feature maps: Average and Concatenation, shown in 2.1 and 2.2 of Fig. 2.

For average aggregation, the output feature takes two feature maps to compute using element-wise average before being fed to the other three components in IVF Block.

For concatenate aggregation, the output feature is a depth stack of two input feature maps before the convolutional layer takes a two-dimension-depth feature map to get the one-dimension-depth feature map. The batch normalization layer and ReLU activation function remain the same as average aggregation for normalizing the input features.

To enhance the examined view with informative information, the output feature map of IVF Block and examined view feature map are combined by element-wise addition. The High-Level Features Blocks (the last half of the backbone) take the enhanced feature maps to learn the high-level information such as abnormalities. Subsequently, we feed it into fully connected layers, followed by a softmax layer, to get the final output binary classification.

This framework's concept is based on how radiologists examine mammograms for diagnosis. Instead of treating two ipsilateral views equally for cancer diagnosis, the model seeks to distinguish one as the primary view and the other as a support view. As shown in Part 2 of Fig. 2, the examined view feature and fused feature play important roles in classifying breast cancer. The examined view is the radiologist's main focus, which is kept the same. On the other hand, the auxiliary view along with examined view is for comparing these two to having more perspectives.

3.2. Malignant Lesions Synthesis Framework

Inherited from the previous successful use of domain adaptation on mammogram classification [27], and mammogram detection tasks [20], we proposed a novel framework to create the natural-looking malignant findings synthesis framework. The framework includes three stages:

1. We first propose a way to select the important region from the benign breast by getting a saliency map from warm-up pre-trained supervised learning. Then, the region with a high-intensity score was replaced by a malignant region that was annotated by radiologists.

2. Secondly, To solve the domain shift issue brought on by breast density or device differences, we conduct the style transfer from the source region to the target region based on Fourier Domain Adaptation [29].

3. Finally, to make the malignant lesions naturally mix with the destination region, we propose a soft contour mask and its inverse to combine source and target regions before pasting to a region of the benign sample.

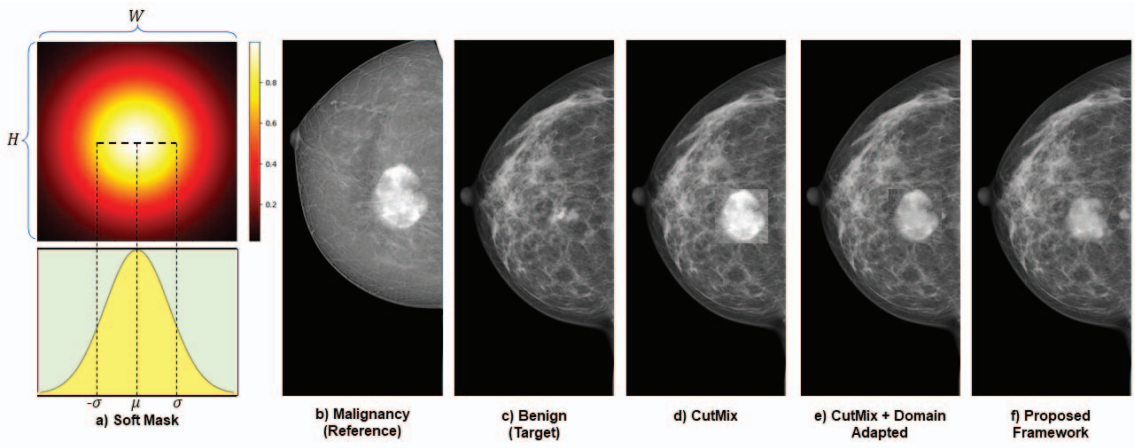

Figure 4. The synthesis mammogram image with various algorithms. a) Soft Mask: following the Gaussian distribution (bottom) to generate the blending masks (top), b) Reference image contains malignancy mass and c) Target image will be added malignancy mass, d) the hard region synthesis CutMix algorithm image, e) the middle smooth region synthesis image with CutMix and Domain Adapted algorithms, f) Our proposed Soft-Adapted Malignancy Synthesis image.

In the region selection part, the supervised training for warming up is conducted before getting a saliency map. Grad-CAM [25] uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest. Mathematically, with given class c, the saliency map from Grad-CAM $L^c_{GC} \in R^{H \times W}$ of height H and width W is obtained by computing the gradient of the class c score and y^c with respect to feature map activations A^k of the convolutional layer (denotes by $\frac{\partial y^c}{\partial A^k}$). To obtain the neuron importance weights (called a^c_k), these gradients are global-average-pooled over the width and height dimensions (indexed by i and j, respectively):

$$a^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}. \tag{1}$$

To obtain L^c_{GC}, we perform a weighted combination of forward activation maps followed by a ReLU:

$$L^c_{GC} = ReLU\left(\sum_k a^c_k A^k\right). \tag{2}$$

Based on radiologists' malignant abnormalities annotations with width W_r and height H_r, coordinates (i, j) are set as top-left corner starting coordinates. We calculate the bottom-right saliency area value with beginning coordinates i, j and the shape of regions H_r, W_r. Region value of class-discriminative localization map from Grad-CAM, called $I_{region}\left(L_{GC}, i, j, H_r, W_r\right)$, is defined as:

$$I_{region}\left(L_{GC}, i, j, H_r, W_r\right) = \sum_{m=i}^{H_r+i-1} \sum_{n=j}^{W_r+j-1} L_{GC}\left(m, n\right). \tag{3}$$

For selecting a wanted region given class c as a pasting destination, we compute the values of regions and find the highest class-discriminative patch as below:

$$I^*_{region} = I_{region}\left(L_{GC}, i^*, j^*, H_r, W_r\right), \tag{4}$$

whereas i^*, j^* are computed by:

$$i^*, j^* = \arg\max_{i,j} I_{region}\left(L_{GC}, i, j, H_r, W_r\right). \tag{5}$$

Using i^*, j^* with H_r, W_r, we can get the patch containing the benign information for mixing. The detailed pseudo-code is described in Algorithm 1 below.

Algorithm 1 High class-discriminative Region Selection

Input: H, W, H_r, W_r.
for $i = 1$ to $H - H_r + 1$ **do**
  for $j = 1$ to $W - W_r + 1$ **do**
    Calculate I_{region} using Eq.3 {Compute accumulative intensity of Region saliency map.}
    if $I^*_{region} < I_{region}$ **then**
      $I^*_{region} \leftarrow I_{region}$ {Update the biggest intensity of region.}
      $(i^*, j^*) \leftarrow (i, j)$ {Update coordinate of the biggest intensity of region.}
    end if
  end for
end for
Return: (i^*, j^*) and I^*_{region}.

Next, in the domain adaptation stage, the domain shift problem between two patches, which brings the different bright fields and device information, could make the noise

for model training. Inspired by FDA [29], the proposed framework conducts spectral transfer, mapping a benign sample to a malignant sample without changing semantic content. Given that F^A, $F^P : R^{H\times W\times 1} \rightarrow R^{H\times W\times 1}$ are the amplitude and phase components of the Fourier transform F of a mammogram patch, we have:

$$F(x)(m,n) = \sum_{h,w} x(h,w) e^{-k2\pi\left(\frac{h}{H}m + \frac{w}{W}n\right)}, \quad (6)$$

where $k^2 = -1$.

With mask M_β contains zero value except for center region with $\beta \in (0,1)$ as follows:

$$M_\beta(h,w) = \Im_{(h,w)\in[-\beta H:\beta H,-\beta W:\beta W]}, \quad (7)$$

$\Im$ indicates an all-ones matrix. As shown in Fig. 3, Benign patch and Malignant patch are $x^s \sim D^s$,$x^t \sim D^t$ respectively, FDA algorithm is shown as:

$$x^{s\rightarrow t} = F^{-1}\left(M_\beta \circ F^A(x^t) + (1 - M_\beta) \circ F^A(x^s), F^P(x^s)\right), \quad (8)$$

where F^{-1} is the inverse Fourier transform mapping spectral information back to 2D-image space. The center (low frequency) part of the amplitude of the source image $F^A(x^s)$ will be transferred in the target style of x^t. This notation only modifies the amplitude component without altering the phase component F^P. Both components of the Fourier transform will be inversed back to a new image $x^{s\rightarrow t}$, whose remaining content of source image x^s but the style of target image x^t.

Finally, the original malignant and domain-adapted benign are used for blending before pasting back to the benign sample. We proposed a novel soft mask and its inverse for mixing two patches. With any image having height H and width W, a soft mask S is defined as $S \in [0,1]^{H\times W}$. Therefore, its inverse soft mask is $(1 - S) \in [0,1]^{H\times W}$. The output image mixing between two images x^s, x^t is formularized as:

$$\overline{x} = \left(S \otimes x^t\right) \oplus \left((1 - S) \otimes x^s\right), \quad (9)$$

whereas, x^s, x^t are the source image (benign patch) and target image (malignant patch) respectively. The label of image $\overline{x}$ is the label of the target image.

The blending masks are generated following the Gaussian distribution. The gradient radial soft mask is the result of the outer product of two one-dimensional Gaussian distributions. It can be seen as:

$$S_W = e^{-\frac{(x-\mu_W)^2}{2\sigma^2}},\ S_H = e^{-\frac{(x-\mu_H)^2}{2\sigma^2}}, \quad (10)$$

whereas μ_W, μ_H, and σ are uniformly sampled from the input images' width W, height H ranges, and its spread in the image space, respectively. A sample of the mask can be seen in Fig. 3 and 4a.

4. Experimental Settings

4.1. Datasets

CMMD. The Chinese Mammography Database (CMMD) [2] includes 5.202 screening mammogram images conducted on 1.775 studies. We trained on 1.172 non-malignant mammograms and 2.728 malignant screening images with 85%:15% ratio splitting on the training set and test set. Furthermore, we employ stratified sampling, resulting in 498 benign and 1157 malignancy ipsilateral view samples on the training set and 88 benign and 205 malignancy ipsilateral view samples on the testing set.

VinDr-Mammo. A large-scale full-field digital mammography dataset [22], which contains 20.000 scans from 5.000 studies of Vietnamese patients. Because of the untrustworthy of BI-RADS 3, the inconsistency between BI-RADS 4 and 5, and the heavy imbalance of BI-RADS 1, we arrange the image-level labels into two classes: Suspicious Benign (BI-RADS 2) and Suspicious Malignancy (BI-RADS 4 and 5). So as the preprocessing on CMMD, there are 2.831 ipsilateral view samples (CC and MLO views on the same breast) that were split into training set (1870 benign and 395 malignancy cases) and testing set (467 benign and 99 malignancy cases). Besides, for the malignant lesions synthesis framework, we use all region-level annotations for making new malignant samples.

4.2. Implementation Details

ResNet family architectures are used for the Feature Extractor part of the framework, including ResNet-18 and ResNet-34. In the data loading part, the images are loaded with a batch size of 32 (two ipsilateral views for each breast with a total of 16 breasts). The model was trained for 200 epochs using SGD optimizer [11] with an initial learning rate 1×10^{-3} and decays by 0.1 after $20, 40, 60$, and 80 epochs. We resized images to the same 800 x 800 for both the training and testing phases. Our work was built on Pytorch version 1.9.1 and trained by using NVIDIA RTX 3090Ti GPU (24GB). We used the Macro F1-Score to evaluate and reduce the imbalance effectiveness in the dataset, which is computed using the arithmetic (unweighted) mean of all the per-class F1 scores. Besides, the Area under the ROC Curve (ROC AUC) is used for measuring the model performance under slightly imbalanced dataset training.

5. Results and Ablation Studies

5.1. Dual Ipsilateral Views Fusion Network

In this section, there are three main approaches we want to test. 1) No Fusion, a single view is fed into the backbone,

Table 1. Quantitative results (%) among our proposed DIVF frameworks, normal fusion frameworks, and no fusion approach

Backbone		**ResNet-18**		**ResNet-34**	
Dataset	Method	F1-Score	AUC-ROC	F1-Score	AUC-ROC
VinDr-Mammo	No Fusion	70.12	68.79	71.48	70.22
	Average Fusion	72.54	74.20	73.25	72.88
	Concatenate Fusion	73.22	70.66	74.63	72.18
	DIVF(Average)	74.00	72.15	74.17	71.67
	DIVF(Concatenate)	**75.34**	**74.24**	**75.98**	**74.86**
CMMD	No Fusion	73.26	76.70	75.52	77.18
	Average Fusion	79.22	79.13	79.97	81.80
	Concatenate Fusion	75.86	77.10	78.12	77.67
	DIVF(Average)	**81.45**	**84.14**	**82.44**	80.92
	DIVF(Concatenate)	77.77	80.42	79.51	**81.97**

no combining of two views CC and MLO in this case. 2) Average Fusion and Concatenate Fusion, there is no skip connection with two examined features and fused features in the Features Fusion Blocks phase. 3) DIVF, contains all components described in Section 3.1. Table 1 shows the testing results of our proposed methods on VinDr-Mammo and CMMD datasets. Our DIVF framework shows a significant improvement compared to the conventional techniques, with a mean of around 5% on VinDr-Mammo and 7% on CMMD. For each method of combining features, the DIVF shows the apparent effectiveness of the feature fusion mechanism for classifying the benign and malignant. Testing on VinDr-Mammo, DIVF Framework with concatenate method achieves the highest Macro F1-score and AUC-ROC on both backbones, 75.98% and 74.86% respectively. Different from VinDr-Mammo, we use average aggregation with the DIVF method seems to be more robust on CMMD. This strategy outperforms the normal fusion or no fusion approaches, which achieved 81.45% on ResNet-18 and 82.44% on ResNet-34 in Macro F1-score evaluation metrics.

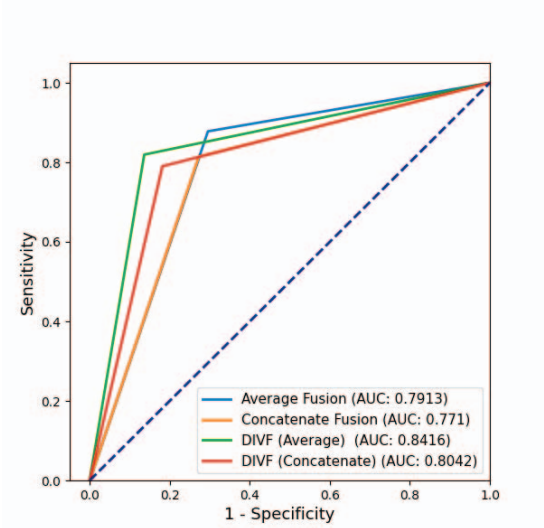

Figure 5. AUC-ROC for benign/malignant classification on CMMD dataset. Testing performance of the average fusion, concatenate fusion, EA average fusion, and EA concatenate fusion

Fig. 5 highlights the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) by plotting the ROC curve for malignant and benign categorization. The DIVF (Average) achieved the best performance with a sensitivity of 87.8% and a specificity of 70.45%, which resulted in 0.8416 of AUC. Continuously, the second high performance is the DIVF (Concatenate), which obtained 80.42%, lower than 3.74% compared with the best one. In contrast, Average Fusion and Concatenate Fusion do not outcome the DIVF version which achieved 79.13% and 77.1%, respectively. This overcome can be explained in the way we support the model by adding the examined features in the features fusion blocks phase of the framework. After going through the IVF block, fused features might lose detailed information because fusion operation tends to generalize the feature in both views. Thus, this alleviates the examined features. Therefore, adding the examined features prevents two problems: solving the vanishing problem in the IVF block and diversifying information.

5.2. Soft-Adapted Malignancy Synthesis Framework

Table 2 shows the ablation studies of our proposed synthesis framework on VinDr-Mammo malignant sample synthesis. As shown in the table, we can see the effect of each element contributing to the final F1 score of our method. The whole framework combined three mechanisms for creating new samples achieves 77.02% on the F1-Score metric. The limitation of the original CutMix seems to be eliminated with Fourier Adaptation and Soft Mask. The new

Table 2. Ablation studies of our proposed Soft-Adapted Malignancy Synthesis Framework on DIVF Concatenate with ResNet-34 on VinDr-Mammo Dataset

	DIVF	Region Selection & CutMix	Fourier Adaptation	Soft Mask	Macro F1-Score
Baseline	✓				75.98
Proposed Methods	✓	✓			76.54
	✓	✓	✓		76.96
	✓	✓		✓	76.78
	✓	✓	✓	✓	**77.32**

samples are no longer containing bad-looking malignant tumors with different color-style and hard contours when conducting copy-and-paste patches. The detailed outputs of each part in our framework are visualized in Fig. 4b-f.

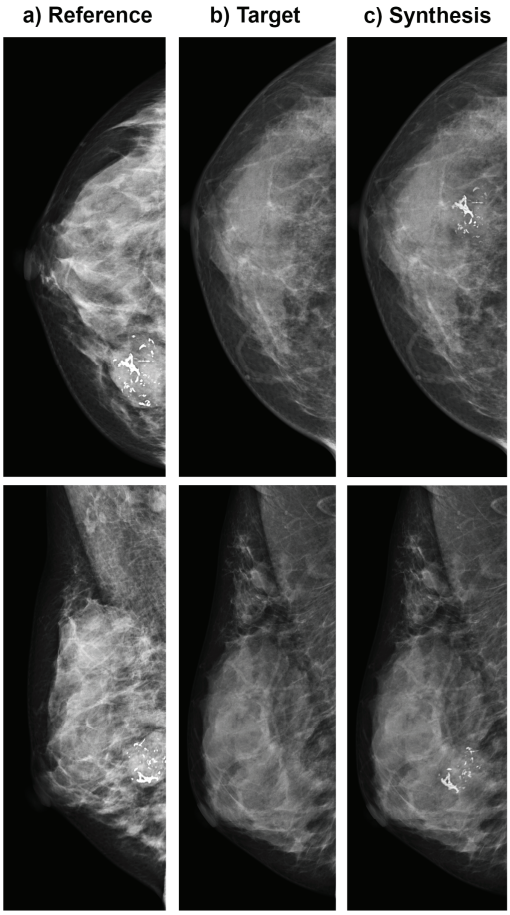

Figure 6. Our framework results on CC (top) and MLO (bottom) views. a) Reference image b) Target image c) Synthesis image.

Fig. 4d shows the synthesis image using the CutMix algorithm. In the replaced region, the cancer mass seems incompatible with the source image in the style. This can cause the unwanted detect edge in CNN sliding filters, thus leading to outlier features and poor representation. This strategy achieved a slight improvement in performance with (+0.56%) compared with the baseline model, ResNet-34 DIVF Concatenate in Table 1. Furthermore, the results also increase a bit (+0.42%) when the Fourier Domain Adaptation method is applied. Fig. 4e proves the improvement with smooth style in the replaced region. However, the suddenly changing pixel value that occurs on the top-left corner of the transformed region does not perfectly make the synthesis image look natural. Afterward, our proposed Soft-Adapted Malignancy Synthesis Framework can alleviate those problems which perfectly adapting the target style to the source image. Fig. 4f and Fig. 6c show natural-looking, yet trustworthy, mammography screening that achieved 77.32% on Macro F1-Score. Those upsampling data shown in Fig. 6c, created by Fig. 6a,b, are reliable for the training stage to handle most of the imbalance mammogram dataset. This framework has shown its robustness on many different types of lesions including Mass, Calcification, Asymmetry, etc.

6. Conclusion

In this work, we proposed a DIVF framework to leverage the ipsilateral multi-view information for classifying cancerous mammograms. The model learns the low-level features separately from two ipsilateral views and conducts feature aggregation for fusion learning on the high-level features. Our model learned low-level features from two ipsilateral views and effectively fused high-level features. The IVF block enhanced the examined view, resulting in improved classification. Additionally, our natural-looking malignant lesions synthesis framework generated reliable samples, leading to state-of-the-art performance and generalizability across two datasets. Our research shows promise for enhancing breast cancer diagnosis and treatment. Future work aims to extend our research to lesion detection and density classification tasks and conduct further statistical analyses to gain deeper insights.

7. Acknowledgement

This paper is partially supported by AI VIETNAM. We thank Integrated MechanoBioSystems Lab (IMBSL) from the Biomedical Engineering Department of National Cheng Kung University for providing the GPU to support the numerical calculations in this paper.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

[2] Hongmin Cai, Jinhua Wang, Tingting Dan, Jiao Li, Zhihao Fan, Weiting Yi, Chunyan Cui, Xinhua Jiang, and Li Li. An online mammography database with biopsy confirmed types. *Scientific Data*, 10(1):123, 2023.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1259–1272, 2018.

[5] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[6] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. Gvcnn: Groupview convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, page 264–272, 2018.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020.

[8] 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.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[11] Nikhil Ketkar and Nikhil Ketkar. Stochastic gradient descent. *Deep learning with Python: A hands-on introduction*, pages 113–132, 2017.

[12] Dimitrios Korkinof, Tobias Rijken, Michael O'Neill, Joseph Yearsley, Hugh Harvey, and Ben Glocker. High-resolution mammogram synthesis using progressive generative adversarial networks. *arXiv preprint arXiv:1807.03401*, 2019.

[13] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*, 2020.

[14] Yuhang Liu, Fandong Zhang, Chaoqi Chen, Siwen Wang, Yizhou Wang, and Yizhou Yu. Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5947–5961, 2021.

[15] Yuhang Liu, Fandong Zhang, Qianyi Zhang, Siwen Wang, Yizhou Wang, and Yizhou Yu. Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3812–3822, 2020.

[16] Yuhang Liu, Zhen Zhou, Shu Zhang, Ling Luo, Qianyi Zhang, Fandong Zhang, Xiuli Li, Yizhou Wang, and Yizhou Yu. From unilateral to bilateral learning: Detecting mammogram masses with contrasted bilateral network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 477–485. Springer, 2019.

[17] Xiangyuan Ma, Jinlong Wang, Xinpeng Zheng, Zhuangsheng Liu, Wansheng Long, Yaqin Zhang, Jun Wei, and Yao Lu. Automated fibroglandular tissue segmentation in breast mri using generative adversarial networks. *Physics in Medicine Biology*, 65(10):105006, 2020.

[18] Rui Man, Ping Yang, and Bowen Xu. Classification of breast cancer histopathological images using discriminative patches screened by generative adversarial networks. *IEEE Access*, 8:155362–155377, 2020.

[19] Hasan Nasir Khan, Ahmad Raza Shahid, Basit Raza, Amir Hanif Dar, and Hani Alquhayz. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, 7:165724–165733, 2019.

[20] Huy T Nguyen, Thinh B Lam, Quan DD Tran, Minh T Nguyen, Dat T Chung, and Vinh Q Dinh. In-context cross-density adaptation on noisy mammogram abnormalities detection. *arXiv preprint arXiv:2306.06893*, 2023.

[21] Huyen TX Nguyen, Sam B Tran, Dung B Nguyen, Hieu H Pham, and Ha Q Nguyen. A novel multi-view deep learning approach for bi-rads and density assessment of mammograms. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2144–2148. IEEE, 2022.

[22] Nguyen H.Q. Pham H.H. et al. Nguyen, H.T. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Sci Data*, 2023.

[23] Vinh Dinh Nguyen, Duc Dung Nguyen, Sang Jun Lee, and Jae Wook Jeon. Local density encoding for robust stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(12):2049–2062, 2014.

[24] Charles R. Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, page 5648–5656, 2016.

[25] 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.

[26] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, page 945–953, 2015.

[27] Yan Wang, Yangqin Feng, Lei Zhang, Zizhou Wang, Qing Lv, and Zhang Yi. Deep adversarial domain adaptation for breast cancer screening from mammograms. *Medical image analysis*, 73:102147, 2021.

[28] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2020.

[29] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020.

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[32] Wenwei Zhao, Runze Wang, Yunliang Qi, Meng Lou, Yiming Wang, Yang Yang, Xiangyu Deng, and Yide Ma. Bascnet: Bilateral adaptive spatial and channel attention network for breast density classification in the mammogram. *Biomedical Signal Processing and Control*, 70:103073, 2021.