HoVer-Trans: Anatomy-Aware HoVer-Transformer for ROI-Free Breast Cancer Diagnosis in Ultrasound Images

Yuhao Mo, Chu Han, *Member, IEEE*, Yu Liu, Min Liu, Zhenwei Shi, Jiatai Lin, Bingchao Zhao, Chunwang Huang, Bingjiang Qiu, Yanfen Cui, Lei Wu, Xipeng Pan, Zeyan Xu, Xiaomei Huang, Zhenhui Li, Zaiyi Liu, Ying Wang, and Changhong Liang

***Abstract*— Ultrasonography is an important routine examination for breast cancer diagnosis, due to its non-invasive, radiation-free and low-cost properties. However, the diagnostic accuracy of breast cancer is still limited due to its inherent limitations. Then, a precise diagnose using breast ultrasound (BUS) image would be significant useful. Many learning-based computer-aided diagnostic methods have been proposed to achieve breast cancer diagnosis/lesion classification. However, most of them require a pre-define region of interest (ROI) and then classify the lesion inside the ROI. Conventional classification backbones, such as VGG16 and ResNet50, can achieve promising classification results with no ROI requirement. But these models lack interpretability, thus restricting their use in clinical practice. In this study, we propose a novel ROI-free model for breast cancer diagnosis in ultrasound images with interpretable feature representations. We leverage the anatomical prior knowledge that malignant and benign tumors have different spatial relationships between different tissue layers, and propose a HoVer-Transformer to formulate this prior knowledge. The proposed HoVer-Trans block extracts the inter- and intra-layer spatial information *horizontally* and *vertically*. We conduct and release an open dataset *GDPH&SYSUCC* for breast cancer diagnosis in BUS. The proposed model is evaluated in three datasets by comparing with four CNN-based models and three vision transformer models via five-fold cross validation. It achieves state-of-the-art classification performance (*GDPH&SYSUCC* AUC: 0.924, ACC: 0.893, Spec: 0.836, Sens: 0.926) with the best model interpretability. In the meanwhile, our proposed model outperforms two senior sonographers on the breast cancer diagnosis when only one BUS image is given (*GDPH&SYSUCC*-AUC ours: 0.924 vs. reader1: 0.825 vs. reader2: 0.820).**

***Index Terms*— Breast cancer diagnosis, transformer, ultrasound, anatomical structure.**

I. Introduction

BREAST cancer is the most commonly diagnosed cancer and the leading cause of cancer death in women globally [1]. Early breast cancer diagnosis can reduce mortality and increase survival rates [2]. Breast ultrasound (BUS) is an important imaging modality for breast cancer diagnosis and screening because it is low-cost, non-invasive, radiation-free, and relatively more sensitive for dense breast tissue [3]. In addition, ultrasound is effective at the differentiation of cysts from solid lesions [4]. Therefore, it is meaningful for breast cancer patients if there exists a precise diagnostic method for BUS, especially for the dense breast patients.

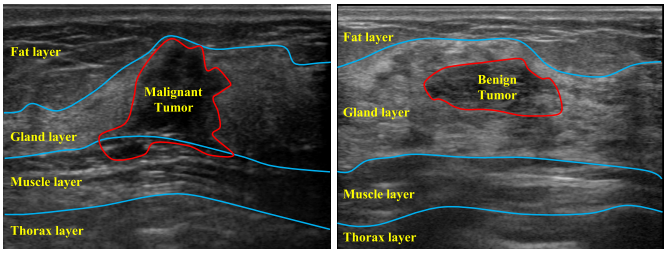
Currently, BUS evaluation generally relies on the subjective evaluation of sonographers. However, the diagnostic accuracy of ultrasound is constrained by the limited number of specialized sonographers. In addition, a high intra- and inter-observer variability exists even among expert sonographers. To overcome such difficulties, a series of computer-aided diagnosis (CAD) systems [5], [6], [7] have been constructed to help sonographers with a more efficient and more precise breast cancer diagnosis. With the recent advancements of deep learning models, the performance of the diagnostic models even outperforms expert sonographers [8] and can reduce the false-positive rate for sonographers [9]. Even though existing models have already achieved outstanding diagnostic performance, most of them are st ill a 'black box', which lacks interpretability. Furthermore, since the open-source data in the community is very limited, existing models either evaluated

Manuscript received 26 October 2022; revised 7 January 2023; accepted 8 January 2023. Date of publication 11 January 2023; date of current version 1 June 2023. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2021B0101420006; in part by the Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application under Grant 2022B1212010011; in part by the National Key Research and Development Program of China under Grant 2021YFF1201003; in part by the National Science Fund for Distinguished Young Scholars under Grant 81925023; in part by the Regional Innovation and Development Joint Fund of National Natural Science Foundation of China under Grant U22A20345; in part by the National Science Foundation for Young Scientists of China under Grant 62102103, Grant 62002082, Grant 82202142, and Grant 82102034; in part by the National Natural Science Foundation of China under Grant 82272084, Grant 82272088, Grant 82271941, and Grant 82071892; in part by the High-level Hospital Construction Project under Grant DFJHBF202105; and in part by the China Postdoctoral Science Foundation under Grant 2021M700897, Grant 2021M690753, and Grant 2022M710843. *(Yuhao Mo, Chu Han, Yu Liu, and Min Liu contributed equally to this work.) (Corresponding authors: Chu Han; Zaiyi Liu; Ying Wang; Changhong Liang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences under Application No. GDRECKY2020-019-01.

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/TMI.2023.3236011



(a) Malignant tumor (a) Benign tumor

Fig. 1. Anatomical structure of breast in ultrasound images. (a) Malignant tumor. (b) Benign tumor.

the performance on a relatively small open dataset (BUSI [10] or UDIAT [11]) or their private datasets. The clinical usability of the diagnostic model should be further assessed.

In this paper, we promote automatic breast cancer diagnosis in ultrasound images in the following two aspects, data resources and methodology. First, we release a large breast lesion classification dataset *GDPH&SYSUCC*, which was collected from two medical centers with 886 benign and 1519 malignant images, for a total of 2405 BUS images. We only provide the whole BUS images without ROI annotations. Then, we propose an interpretable breast cancer diagnosis model for BUS. We find that the sequential data analysis nature of the transformer perfectly fits the anatomical prior of the breast ultrasound image. As shown in Fig. 1, there are four layers from top to bottom, including subcutaneous fat layer, gland layer, muscle layer and thorax layer. Malignant tumors always start from the gland layer (horizontally) and invade the deeper layers (vertically). Benign breast tumors typically originated in the glandular tissue and destructed the gland continuity (horizontally). Therefore, we design an anatomy-aware model, called HoVer-Transformer (in short HoVer-Trans), which considers the prior knowledge of the anatomical structure in BUS. We propose a HoVer-Trans block to extract the inter-layer spatial information *horizontally* and the intra-layer spatial information *vertically*. In HoVer-Trans, we introduce convolutional layers to joint two adjacent transformer stages to fuse the horizontal and vertical image features and to introduce inductive bias.

The proposed HoVer-Trans is evaluated by extensive experiments, including comparisons with state-of-the-art (SOTA) methods in several datasets, model interpretability and ablation studies. HoVer-Trans achieves comparable quantitative performance in all the datasets. The visualization heatmaps also demonstrate that HoVer-Trans is able to pay attention to the malignant lesion boundary (invasive margin), which proves the horizontal and vertical design successfully learned anatomical prior knowledge. Ablation studies demonstrate the effectiveness of each specific technical design. We also compare our proposed model with two senior sonographers. HoVer-Trans outperforms both sonographers in the entire dataset and Breast Imaging-Reporting and Data System(BI-RADS) subgroup analysis under the same condition. The main contributions of this paper are summarized as follows.

- We release a new breast cancer classification dataset, GDPH&SYSUCC which is the largest open dataset in this field.
- We propose an anatomy-aware model HoVer-Trans to fully automatically classify the breast lesion and achieve comparable performance compared with six baseline models.
- HoVer-Trans is able to provide the interpretable evidence to support the decision of the model.

II. Related Works

Deep learning techniques have dominated almost all the medical imaging modalities, including MRI [12], CT [13], histopathological images [14] and etc. In this section, we summarize the previous research works on breast cancer diagnosis in ultrasound images [15] and the transformer-based medical image classification models [16].

A. Breast Cancer Diagnosis in Ultrasound Images

Ultrasonography is one of the most common and non-invasive imaging modalities for breast cancer screening and diagnosis. Precisely detecting and diagnosing malignant tumors allows early intervention to reduce mortality. Therefore, CAD algorithms haven been designed to automatically and objectively evaluate breast ultrasonography. With the recent advances in deep learning [17], researchers started to solve various clinical prediction applications in a data-driven manner and achieved outstanding performances in breast cancer diagnosis, such as lesion classification [4], [18], axillary lymph node status prediction [19], sentinel lymph node status prediction [20] and even molecular status prediction [21].

Currently, deep learning-based models have already dominated the breast lesion classification. Byra et al. [22] transferred the model pre-trained on ImageNet to fine-tune the breast mass classification model, which is a popular way for small- or mid-sized data. Some researchers [23] ensembled the deep features from multiple classification architectures and applied machine learning classifiers for breast ultrasonography image classification. Zhuang et al. [24] proposed an image decomposition and enhancement method to enrich the information of the ultrasound image. Qian et al. [4] aggregated the multimodal ultrasound images for an explainable prediction to support the clinical decision-making of the sonographers and increase the confidence levels of the decision. Cui et al. [25] proposed an FMRNet to fuse combined tumoral, intratumoral and peritumoral regions to represent the whole tumor heterogeneous. Zhang et al. [26] incorporated both classifying breast tumors and interpretable morphological BI-RADS descriptors for BUS images into the classification task. Zhang et al. [27] converted BUS images to feature maps with BI-RADS features. Feature maps were used to classify breast tumors under semi-supervised learning and to reconstruct feature maps guided by lesion classification under unsupervised learning. Di et al. [28] introduced a saliency-guided approach to differentiate the foreground and background regions by two separated branches. A hierarchical feature aggregation branch was proposed to fuse the features from both branches and make the inference. Qi et al. [29] designed two identical CNN backbones to identify the malignant tumor and solid nodule separately. The class activation maps generated from two CNN

backbones were used to guide each other. They validated the proposed model in a large dataset with 8145 breast ultrasonography images. Unfortunately, the dataset in this paper is private.

For the breast lesion classification, even though existing models have already achieved comparable performance with sonographers, it is still worth to keep discovering the potential values of deep learning models from different perspectives. People now pay more attention to clinical usability other than only considering the accuracy. The clinical usability reflects in the following factors. (1) *Accuracy*: Whether the model prediction results are more accurate than sonographers. (2) *Interpretability*: Whether the model can provide any sonographic symptom or evidence to support the decision. (3) *Model Convenience*: Whether the model is fully automated without any user input, like manual segmentation or predefined ROIs.

B. Transformer-Based Medical Image Classification

Transformer [30] is originally designed for natural language processing. It has been widely used in sequential data analysis thanks to the elegant self-attention mechanism [31]. The invention of the vision transformer (ViT) [32] is leading the transformer-based models toward the computer vision applications by cropping the image into several small tiles (visual words). Swim transformer [33] introduces multi-scale information like a CNN model does by a hierarchical structure and shifted windows. Sooner, various transformer models [34] were proposed for medical image classification, such as COVID-VIT for CT chest COVID-19 classification [35], TransMIL for pathology image classification [36], MIL-VT for fundus image classification [37] and etc.

Instead of designing a complex black-box model for breast cancer prediction, we intend to take model interpretability, generalizability and convenience into consideration. By formulating the anatomical prior knowledge into the transformer model design, the proposed model demonstrates superior predictive ability while can provide interpretable features to support the decisions.

III. Methodology

The anatomical structures of the breast can be observed in the ultrasound images. The malignant tumor and the benign tumor demonstrate different spatial relationship between the lesion and different anatomical layers. By leveraging this prior knowledge, we propose an anatomy-aware model for fully automatic breast cancer diagnosis in ultrasound images. In this section, we demonstrate the methodology of the proposed model, shown in Fig. 2. First, we introduce the key idea of the anatomy-aware formulation in Sec. III-A. Based on this idea, the HoVer-Trans stage is proposed in Sec. III-B. Next, we define the overall network structure of the proposed model in Sec. III-C. Sec. III-D shows the implementation details.

A. Anatomy-Aware Formulation

According to the ultrasound imaging principles and the anatomical structure of the breast, different breast tissues form different layers clearly in the ultrasound images, as shown in Fig. 1. The size, location and morphological appearance of the lesion and the spatial relationship with different layers determine the malignancy of the lesion. Conventional CNN models are good at extracting representative local features but show less effective spatial relationship representation ability. That is the reason why most of the existing breast cancer diagnosis algorithms in ultrasound images need a pre-defined ROI of the lesion to remove the redundant area and let the CNN model classify the ROI. The self-attention nature of the transformer introduces strong spatial relationships of each visual word, as shown in Fig. 3 (a). To further exploit the intra-layer and inter-layer spatial correlations in BUS, we formulate the problem by transforming the square-shape visual words into horizontal and vertical strips to bring the anatomical prior knowledge into the model, as shown in Fig. 3 (b).

B. HoVer-Trans Stage

Since our proposed model is constructed on top of the ViT structure, we first briefly introduce ViT at the beginning of this part. And then we show how the proposed HoVer-Trans stage is constructed.

1) Vision Transformer: Vision transformer (ViT) [32] is the first to bring the most popular technique in natural language processing into the computer vision world. It tessellates the input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$ and regards them as the visual words (tokens), where (H, W, C) and (P, P, C) are the resolution with channels of the input image and the patches, respectively. N is the number of patches. For each visual word, they transform the 2D image into a 1D vector, called patch embedding. Multi-head self-attention mechanism builds spatial correlations across different tokens. The formulation of ViT is shown as:

$$\begin{aligned}\mathbf{z}_0 &= [\mathbf{x}_{class}; \mathbf{x}_p^1\mathbf{E}, \mathbf{x}_p^2\mathbf{E}, \ldots, \mathbf{x}_p^N\mathbf{E},] + \mathbf{E}_{pos}, \\ &\qquad \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \qquad (1)\\ \mathbf{z}'_l &= \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \\ \mathbf{z}_l &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \\ \mathbf{y} &= \mathrm{LN}(\mathbf{z}_L^0) \qquad (2)\end{aligned}$$

where $\mathbf{E}$, $\mathbf{E}_{pos}$, MSA, MLP and LN denote trainable linear projection, position embedding, multi-head self-attention module, multi-layer perceptron module and layer norm, respectively.

In our proposed HoVer-Trans, we use several ViT blocks with exactly the same structure without class embedding to construct the HoVer-Trans block. Thus, we denote all the ViT blocks in the following as $\texttt{Trans}(\cdot)$.

2) Embedding: To formulate the anatomical prior knowledge into the transformer model, we introduce additional two embedding ways shown in Fig. 2 (a), following the idea presented in Sec. III-A. Given the input BUS image $I \in \mathbb{R}^{H\times W\times 3}$, we first apply a convolutional stem [38] to downsample the input image by four $I' \in \mathbb{R}^{H/4\times W/4\times C}$ and introduce early inductive bias. And then, patch embedding, horizontal strip embedding and vertical strip embedding are processed before feeding them into the model. Patch embedding cuts I' into $N \times N$ patches $\mathbf{x}_p^{(r,c)}$, where r and c denote

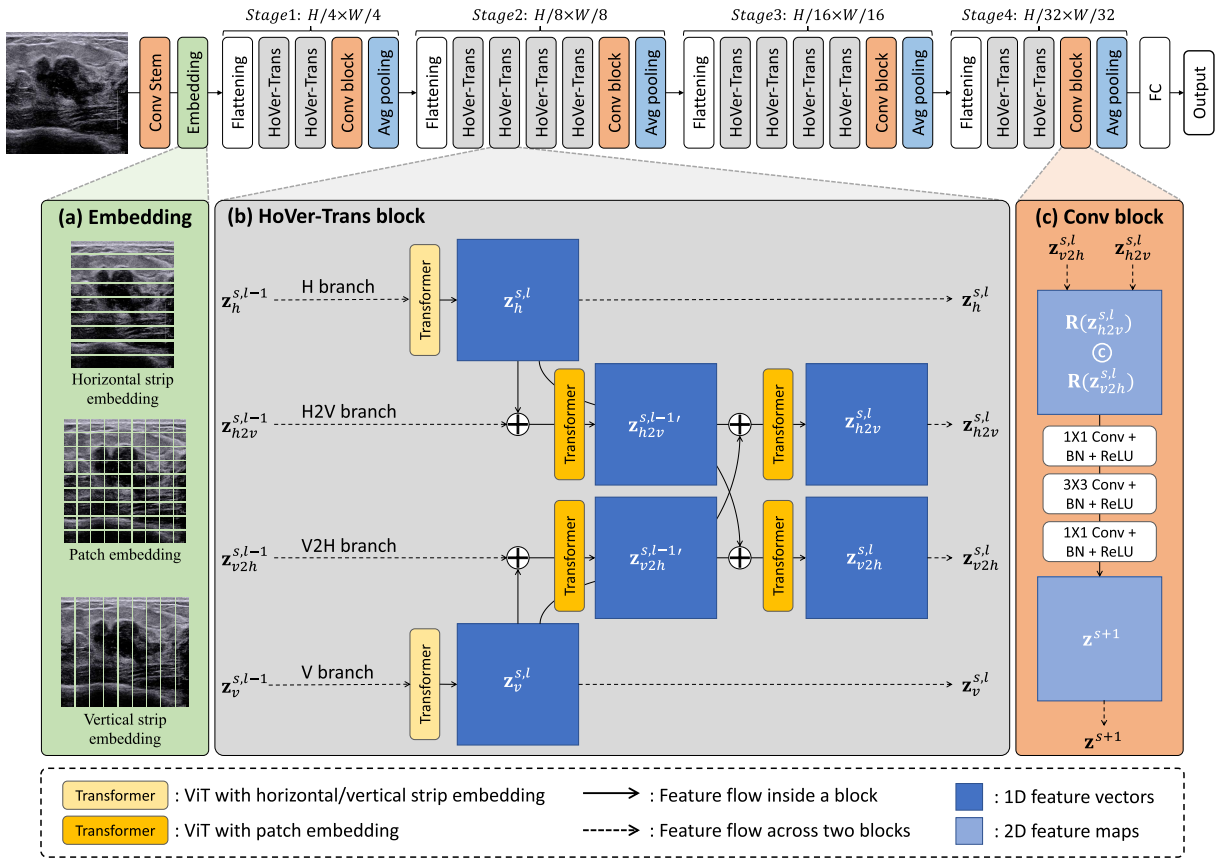


Fig. 2. Network architecture of the proposed model. It contains four stages, each of which consists of a flattening operation, several HoVer-Trans blocks, a convolutional block and a pooling layer. (a) Three embedding ways of horizontal strip embedding, patch embedding and vertical strip embedding. (b) HoVer-Trans formulates the anatomical prior knowledge in breast ultrasound images, which is designed to extract the intra-layer and inter-layer relationships of the anatomical layers in the breast. It consists of four branches. The horizontal branch and the vertical branch are designed to extract the inter-layer and intra-layer relationships respectively. H2V and V2H branches are introduced to fuse the horizontal and vertical features. The output features from each branch in the HoVer-Trans block will be regarded as the input features of the next HoVer-Trans block. (c) Conv block is applied to connect two adjacent stages and to introduce inductive bias.

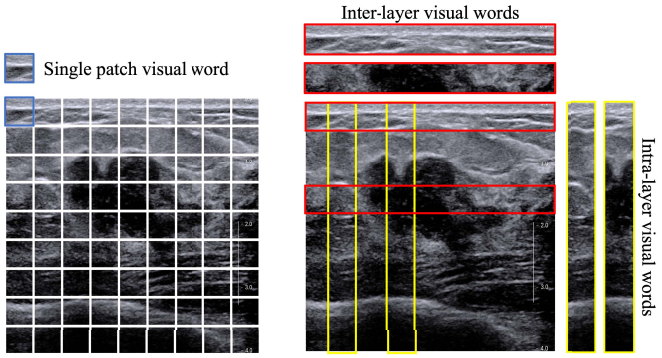


Fig. 3. Comparison of two embedding ways. (a) ViT tessellates the image into several 16×16 patches (visual words). (b) We formulate the anatomical structure of breast by the inter-layer visual words (horizontal strips) and the intra-layer visual words (vertical strips).

the indices of the row and column. After a flattening operation, we get a group of 1D vectors $\mathbf{z}_p$.

$$\mathbf{z}_p = \{\mathbf{x}_p^{(r,c)} | \mathbf{x} \in \mathbb{R}^{H/4N \times W/4N \times C}\}, \quad r, c = 1, \ldots, N \tag{3}$$

Horizontal strip embedding is introduced to represent the visual words of the same anatomical layer with M strips, defined as:

$$\mathbf{z}_h = \{\mathbf{x}_h^{(r)} | \mathbf{x} \in \mathbb{R}^{H/4M \times W/4 \times C}\}, \quad r = 1, \ldots, M \tag{4}$$

Vertical strip embedding is introduced to represent the visual words across anatomical layers with M strips, defined as:

$$\mathbf{z}_v = \{\mathbf{x}_v^{(c)} | \mathbf{x} \in \mathbb{R}^{H/4 \times W/4M \times C}\}, \quad c = 1, \ldots, M \tag{5}$$

3) HoVer-Trans Block: The architecture of the HoVer-Trans block is depicted in Fig. 2 (b). We design a symmetry structure with four branches in one HoVer-Trans block, H branch (horizontal), V branch (vertical), H2V branch (horizontal to vertical) and V2H branch (vertical to horizontal).

Let us define the features at the l-th block in the s-th stage as $\mathbf{z}_{\{h,v,h2v,v2h\}}^{s,l}$. HoVer-Trans block takes the outputs from the previous block and generates the features for the next block, defined as:

$$\{\mathbf{z}_h^{s,l}, \mathbf{z}_v^{s,l}, \mathbf{z}_{h2v}^{s,l}, \mathbf{z}_{v2h}^{s,l}\} = f(\mathbf{z}_h^{s,l-1}, \mathbf{z}_v^{s,l-1}, \mathbf{z}_{h2v}^{s,l-1}, \mathbf{z}_{v2h}^{s,l-1}) \tag{6}$$

where $f(\cdot)$ denotes the HoVer-Trans block. The inputs of four branches in the first HoVer-Trans block (when $l = 1$) are equivalent to the features from the previous HoVer-Trans stage $\mathbf{z}^{s-1}$.

$$\mathbf{z}_{\{h,v,h2v,v2h\}}^{s,1} = \mathbf{z}^{s-1} \tag{7}$$

H and V branches are two auxiliary branches to extract the inter-layer and intra-layer spatial correlations, with two identical H and V branches with horizontal strip embedding (defined in Eq. 4) and vertical strip embedding (defined in Eq. 5), respectively.

$$\mathbf{z}_h^{s,l} = \texttt{Trans}(\mathbf{z}_h^{s,l-1}) \tag{8}$$
$$\mathbf{z}_v^{s,l} = \texttt{Trans}(\mathbf{z}_v^{s,l-1}) \tag{9}$$

The anatomy-aware spatial features $\mathbf{z}_h^{s,l}$ and $\mathbf{z}_v^{s,l}$ will be passed into the next two main branches (H2V and V2H). They are also regarded as the inputs of the next HoVer-Trans block.

H2V and V2H branches are served as the main feature extraction branches which fuse the features from two auxiliary branches (H and V). For example in the H2V branch, the horizontal features $\mathbf{z}_h^{s,l}$ are added to the features from the previous HoVer-Trans block $\mathbf{z}_{h2v}^{s,l-1}$. After a transformer encoder, the vertical features $\mathbf{z}_v^{s,l}$ are added behind. The V2H branch is the mirror of the H2V branch.

$$\mathbf{z}_{h2v}^{s,l} = \texttt{Trans}(\texttt{Trans}(\mathbf{z}_h^{s,l} + \mathbf{z}_{h2v}^{s,l-1}) + \mathbf{z}_v^{s,l}) \tag{10}$$
$$\mathbf{z}_{v2h}^{s,l} = \texttt{Trans}(\texttt{Trans}(\mathbf{z}_v^{s,l} + \mathbf{z}_{v2h}^{s,l-1}) + \mathbf{z}_h^{s,l}) \tag{11}$$

The output features $\mathbf{z}_{h2v}^{s,l}$ and $\mathbf{z}_{v2h}^{s,l}$ will be passed into the next block. Note that, for the last HoVer-Trans Block in each stage, the features will be passed into a Conv Block, described as follows.

4) Conv Block: The transformer is good at processing sequential data and extracting spatial correlations. But it lacks inductive bias. In order to leverage the strength of both transformer and CNN, we introduce a convolutional block (Fig. 2 (c)) after the last HoVer-Trans block at each stage to fuse the H2V and V2H features and introduce the inductive bias.

$$\mathbf{z}^{s+1} = \texttt{Conv}(\mathbf{z}_{h2v}^{s,l}, \mathbf{z}_{v2h}^{s,l}) \tag{12}$$

As shown in Fig. 2 (c), the Conv block takes the output 1D feature vectors $\mathbf{z}_{h2v}^{s,l}$ and $\mathbf{z}_{v2h}^{s,l}$ from two main transformer branches as the inputs. These two feature vectors are reshaped to 2D feature maps and concatenated together. After three convolutional layers, the Conv block outputs the 2D feature maps $\mathbf{z}^{s+1}$ for the Stage. Note that, the settings of three convolutional layers are shown in Fig. 2. The channel numbers of each convolutional layer in four stages are $S1 : \{8, 16, 8\}$, $S2 : \{16, 32, 16\}$, $S3 : \{32, 64, 32\}$ and $S4 : \{64, 128, 64\}$ respectively.

C. Overall Architecture

The overall structure of our model is shown in Fig 2. The model consists of four *stage* modules. Each stage module consists of several HoVer-Trans blocks, one Conv block and one pooling layer.

Given a BUS image $I \in \mathbb{R}^{H\times W\times 3}$, we first apply a convolutional stem for early visual processing. Comparing to the patchy stem of original ViT, introducing early inductive bias by an early convolutional stem [38] can improve the optimization stability and the model performance. Then the dimension of the input image I is reduce to $H/4 \times W/4 \times C$ after the convolutional stem, where $C = 4$ in practice. The sizes of the feature maps in the next three stages are $H/8 \times W/8 \times 2C$, $H/16\times W/16\times 4C$, $H/32\times W/32\times 8C$, which is similar to the structure of the traditional convolutional neural network [39], [40]. To fuse the horizontal and vertical information, a Conv block is introduced to connect two adjacent stages. So the input of each stage is a 2D image or 2D feature maps. Embedding or flattening will be introduced to fit the input of the transformer. In the last stage, the fully connected layer is applied for inference. The model is optimized by the cross-entropy loss.

D. Implementation Details

We use Python3.6 and PyTorch1.8 to implement all the models. All the experiments are run with an 11 GB NVIDIA GeForce RTX 2080Ti GPU. We build our model with embedding dimensions of each stage of $\{4, 8, 16, 32\}$, and HoVer-Trans block numbers of each stage are $\{2, 4, 4, 2\}$. Head numbers of transformer block in each stage are $\{2, 4, 8, 16\}$. We train for 250 epochs with the AdamW [41] optimizer, a batch size of 32, weight decay of 0.1, 10 warm-up epochs and an initial learning rate of 0.0001 with a cosine decay learning rate scheduler. The augmentation strategy includes blurring, noise, horizontal flip, brightness and contrast. Because the order of the tissue layers is fixed, we do not use vertical flip data augmentation. All the images will be resized to 256×256. The source code is avaiable at https://github.com/yuhaomo/HoVerTrans.

IV. Datasets

In this paper, we use three datasets to evaluate the diagnostic performance of our model, two of which are the public datasets and one is our constructed dataset.

UDIAT[1]*:* The first public dataset is a small dataset, named UDIAT [11], which contains a total of 163 BUS images with 109 BUS images of benign lesions and 54 BUS images of malignant lesions. All the images were collected from the UDIAT Diagnostic Centre of the Parc Tauli Corporation, Sabadell, Spain. The average size of the images is 760×570 pixels and the range from 307×233 to 791×641.

BUSI[2]*:* The second dataset, BUSI [10], consists of a total number of 780 BUS images from the Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo, Egypt. BUSI dataset includes 210 images with benign lesions, 437 images with malignant lesions, and 133 normal BUS images without lesions. Furthermore, each image has the pixel-level ground truth of the lesion. In this paper, since we only differentiate the malignant and benign lesions, normal BUS images are excluded. 647 images are finally utilized with the average resolution of 608×494 pixels and the range from 190×335 to 916×683.

GDPH&SYSUCC[3]*:* In this study, we also construct a publicly available dataset of BUS images for breast cancer diagnosis. The BUS images came from two medical centers. 1) the Department of Ultrasound, Guangdong Provincial

[1] https://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php
[2] https://scholar.cu.edu.eg/?q=afahmy/pages/dataset
[3] https://github.com/yuhaomo/HoVerTrans

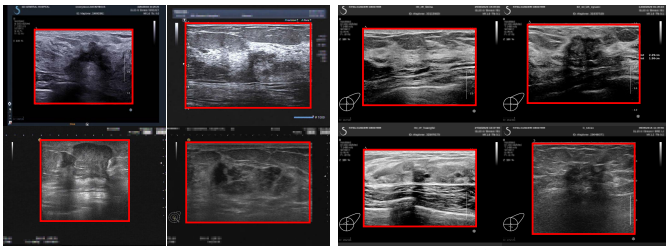


Fig. 4. Some examples of breast ultrasound images in two medical centers, (a) GDPH and (b) SYSUCC. The red boxes are the foreground images we extract.

TABLE I
DISTRIBUTION OF MASS ACCORDING TO BI-RADS STRATIFICATION. BI-RADS STRATIFICATION ARE CATEGORIZED AS INTO 6 CLASSES. CLASS 2 TO 5, INCLUDING 3 SUB-CATEGORIES 4A, 4B AND 4C

BI-RADS	Benign	Malignant	Total
2	66	0	66
3	610	101	711
4A	162	659	821
4B	38	495	533
4C	10	193	203
5	0	71	71
Total	886	1519	2405

People's Hospital, Guangzhou, Guangdong, China (GDPH). 2) the Department of Ultrasound, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China (SYSUCC). We exported the images and their corresponding BI-RADS scores from the picture archiving and communication system (PACS). The ultrasound images were acquired from following equipments, including Hitachi Ascendus (Japan), Mindray DC-80 (China), Toshiba Aplio 500 (Japan) and Supersonic Aixplorer (France). All the images were labeled as benign or malignant according to the pathology report after the biopsy or surgery was performed. The dataset consists of 1519 malignant BUS images from 676 patients and 886 benign BUS images from 526 patients, for a total of 2405 BUS images. The average size of the images is 844×627 and the range from 278×215 to 1280×800. Fig. 4 shows some examples of BUS images in our dataset. To protect the privacy of the patients, we mosaic the personal information. The distribution of the BI-RADS scores and the statistics of the malignant tumors and the benign tumors are shown in Table I.

We extract the image data from the original DICOM format files of BUS and save them in PNG format. To exclude the UI regions, we extract the foreground images by applying a rectangle detection algorithm provided by the OpenCV library, and manually check all the images to ensure the completeness of the foreground images.

V. EXPERIMENTAL RESULTS

In this section, we first introduce the experimental setting in Sec. V-A, including the evaluation metrics and the competitors. Then we show the quantitative results and the heatmap visualizations of three datasets in Sec. V-B. In Sec. V-C, we compare our proposed model with two senior sonographers. Ablation studies have been conducted in Sec. V-D. Heatmap visualizations of the model are shown in all the experiments to evaluate the interpretability.

A. Experimental Setting

Evaluation Metrics: To comprehensively evaluate the performance of the proposed model, we introduce the following metrics, including area under the ROC curve (AUC), accuracy (ACC), specificity, precision, recall and F1 score.

Competitors: In this paper, we compare our proposed model with six SOTA models, including two most popular CNN-based classification models ResNet50 [39] and VGG16 [40], three vision transformer models ViT [32], TNT-s [42] and Swin-B [33] and two CNN-based models tailored for breast cancer diagnosis in ultrasound images BVA-Net (JBHI2020) [43] and MsGoF (MICCAI2020) [18]. Although BVA-Net and MsGoF are both designed for breast cancer diagnosis, they are still different from our proposed model. Both BVA-Net and MsGoF do not process the entire ultrasound images. These two models have to first pre-define a region of interest (ROI) of the mass, and then classify the malignancy of the corresponding lesion. Furthermore, BVA-Net includes additional BI-RADS scores in the training phase. So in the existing two public datasets UDIAT and BUSI, the quantitative performance of these two models is directly copied from the corresponding papers. In the GDPH&SYSUCC dataset, we compare our proposed model with the other four image classification baseline models without ROIs of the lesions. We also implement the model from BVA-Net [43] and test it in the GDPH&SYSUCC dataset. Since BVA-Net requires the pre-defined ROI of the lesion, we invited a sonographer to label the bounding box of each lesion for this approach.

B. Comparisons With SOTA Approaches

Table II shows the quantitative results in three datasets. In the UDIAT dataset, since MsGoF and BVA-Net classify the lesion within the ROIs, it can alleviate the underfitting problem when the dataset only contains 163 images by removing the regions without lesions. Among the other five models, VGG16 and our model achieve promising classification results even with such a small dataset. In the BUSI dataset (647 images), the classification performance of our proposed model achieves the best ACC of 0.855, the precision of 0.876, the recall of 0.867 and the F1-score of 0.872. A larger dataset with more training samples allows the neural network models to learn better feature representation. The performance of our model in the BUSI dataset is comparable to the performance of the ROI-based model BVA-Net.

Most of the existing approaches, including MsGoF and BVA-Net, solve the BUS classification problem by a two-step approach. Pre-defined ROIs can remove the regions without tumors, which might be good for the small dataset. However, when we have enough data, it is hard to say whether removing these regions do more good than harm. Because it also loses the spatial information of the tumor in the breast, which is also an important clue for breast cancer diagnosis. In our dataset GDPH&SYSUCC (2405 images), the proposed HoVer-Trans achieves the best classification performance on AUC, ACC, precision, recall and F1-score. It outperforms all the baseline models, including BVA-Net. When with enough training data, the advantage of the anatomy-aware design is demonstrated.

TABLE II

QUANTITATIVE COMPARISONS WITH SOTA APPROACHES IN THREE DATASETS. THE LAST COLUMN SHOWS THE P-VALUE OF DELONG'S TEST BETWEEN THE AUC OF EACH BASELINE MODEL AND HOVER-TRANS MODEL. P-VALUES LESS THAN 0.05 ARE MARKED AS *, P-VALUES LESS THAN 0.01 ARE MARKED AS **, AND P-VALUES LESS THAN 0.001 ARE MARKED AS ***. MODEL WITH † MEANS THEY REQUIRE A PRE-DEFINED ROI

	AUC	ACC	Specificity	Precision	Recall	F1-score	p-value
UDIAT							
ResNet50	0.778±0.059	0.743±0.073	**0.899±0.118**	0.676±0.146	0.426±0.256	0.523±0.120	***
VGG16	**0.786±0.073**	0.756±0.123	0.800±0.106	0.650±0.120	**0.672±0.183**	**0.661±0.107**	***
ViT	0.740±0.140	0.701±0.300	0.880±0.147	0.606±0.240	0.364±0.132	0.455±0.223	***
TNT-s	0.627±0.082	0.626±0.089	0.752±0.150	0.426±0.276	0.370±0.241	0.396±0.160	***
Swin-B	0.760±0.141	0.761±0.078	0.895±0.119	0.697±0.238	0.495±0.210	0.547±0.186	***
Ours	0.781±0.118	**0.774±0.061**	0.889±0.128	**0.714±0.214**	0.545±0.232	0.619±0.099	-
MsGoF†	0.939±0.031	0.909±0.032	0.927±0.106	0.900±0.044	-	-	-
BVA-Net†	0.870	0.859	0.685	0.945	0.840	-	-
BUSI							
ResNet50	0.877±0.034	0.818±0.039	0.883±0.030	0.738±0.049	0.682±0.079	0.709±0.062	***
VGG16	**0.898±0.037**	0.832±0.041	0.778±0.056	0.873±0.054	0.862±0.096	0.867±0.063	***
ViT	0.834±0.062	0.811±0.052	**0.922±0.070**	0.781±0.146	0.579±0.032	0.665±0.094	***
TNT-s	0.852±0.015	0.812±0.032	0.908±0.035	0.763±0.057	0.611±0.104	0.679±0.057	***
Swin-B	0.858±0.024	0.818±0.026	0.880±0.045	0.736±0.063	0.694±0.084	0.710±0.044	***
Ours	0.865±0.066	**0.855±0.050**	0.838±0.053	**0.876±0.062**	**0.867±0.115**	**0.872±0.080**	-
BVA-Net†	0.889	0.843	0.758	0.883	0.751	-	-
GDPH&SYSUCC							
ResNet50	0.886±0.014	0.832±0.014	0.732±0.033	0.851±0.015	0.890±0.013	0.870±0.010	**
VGG16	0.919±0.006	0.864±0.004	**0.892±0.010**	0.811±0.009	0.814±0.007	0.813±0.003	*
ViT	0.806±0.021	0.734±0.029	0.694±0.053	0.809±0.047	0.758±0.021	0.782±0.028	***
TNT-s	0.853±0.010	0.781±0.015	0.618±0.059	0.793±0.050	0.879±0.028	0.834±0.015	***
Swin-B	0.886±0.024	0.824±0.025	0.744±0.035	0.853±0.019	0.871±0.029	0.865±0.019	**
Ours	**0.924±0.016**	**0.893±0.021**	0.836±0.038	**0.906±0.023**	**0.926±0.019**	**0.916±0.019**	-
BVA-Net†	0.856±0.009	0.811±0.022	0.859±0.038	0.824±0.022	0.891±0.028	0.856±0.020	***

Considering both the tumor region and the surrounding area can introduce more useful information for diagnosis. And the horizontal and vertical formulation combined with the vision transformer allows the model to analyze not only the tumor region itself but also the spatial relationship between the lesion and the different layers in the breast.

Besides the quantitative results, we also demonstrate where the models focus by showing the heatmaps to evaluate the interpretability of the models (in the GDPH&SYSUCC dataset). Fig. 5 shows the heatmaps overlaid on the original images. We point out the lesions on the original image by white arrows in Fig. 5 (a) for better visualization. As can be seen, three transformer-based models in Fig. 5 (e)-(g) cannot focus on the lesions and contains false-positive highlighted areas, which lead to poor specificity (TNT-s: 0.618, ViT: 0.694, Swin-B: 0.744) shown in Table II. Visualization of two CNN-based models ResNet50 and VGG16 are shown in Fig. 5 (c)&(d). ResNet50 also has the same problem with transformer-based models. VGG16 can achieve better visualization results compared with the previous three models. But it also pays attention to the dark areas caused by signal attenuation. Our proposed model in Fig. 5 (b) shows the best visualization results with more accurate lesion locations and more focused attention, thus achieving the best F1-score of 0.916. Fig. 6 demonstrates more heatmaps of the HoVer-Trans model in GDPH&SYSUCC.

C. HoVer-Trans Vs. Sonographers

A reader study is conducted to compare the classification performance of HoVer-Trans with that of the experienced sonographers (YL and YW, at least ten years experience), one from the Department of Ultrasound, Guangdong Provincial People's Hospital and the other from the Department of Medical Ultrasonics, The First Affiliated Hospital of Guangzhou Medical University. The entire GDPH&SYSUCC dataset is presented to the readers in a random order to assess the benignity or malignancy of the BUS lesion. To compare the entire dataset, we aggregate the model prediction results of each fold in the five-fold cross-validation.

TABLE III

HOVER-TRANS VS. SONOGRAPHERS IN THE ENTIRE GDPH&SYSUCC DATASET AND TWO SUBGROUPS. THE LAST COLUMN SHOWS THE P-VALUE OF DELONG'S TEST BETWEEN THE AUC OF EACH READER AND HOVER-TRANS. P-VALUES LESS THAN 0.001 ARE MARKED AS ***

	AUC	ACC	Spec	Prec	Rec	F1	p-value
GDPH&SYSUCC							
reader1	0.825	0.836	0.781	0.872	0.868	0.870	***
reader2	0.820	0.838	0.751	0.859	0.889	0.874	***
Ours	**0.924**	**0.893**	**0.836**	**0.906**	**0.926**	**0.916**	-
BI-RADS 2-3							
reader1	0.503	0.867	**0.996**	0.250	0.099	0.019	***
reader2	0.669	0.828	0.883	0.368	0.455	0.407	***
Ours	**0.886**	**0.870**	0.879	**0.500**	**0.812**	**0.619**	-
BI-RADS 4-5							
reader1	0.510	0.821	0.090	0.873	0.929	0.901	***
reader2	0.622	0.843	0.324	0.902	0.920	0.911	***
Ours	**0.891**	**0.904**	**0.700**	**0.955**	**0.934**	**0.944**	-

The uppermost part in Table III shows the comparison between our model and two readers in all the evaluation metrics in the entire GDPH&SYSUCC dataset. Experimental

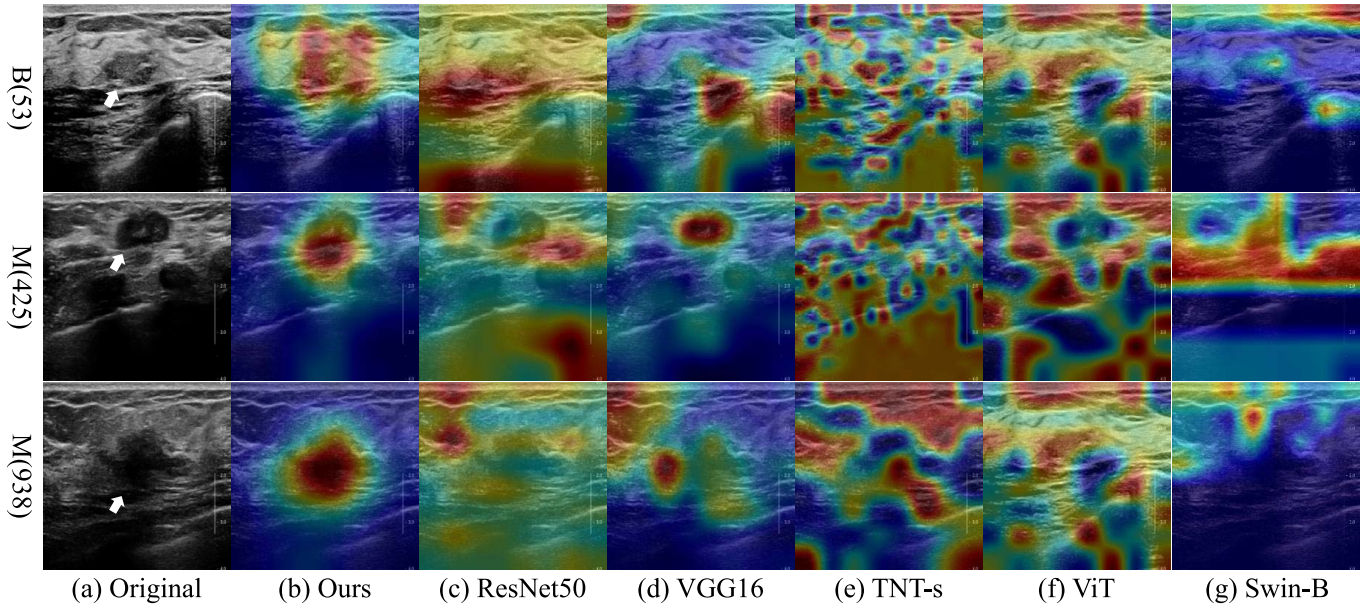


Fig. 5. The heatmaps of different models (GDPH&SYSUCC). We overlay the heatmaps generated by feature maps on the original BUS images. (a) are the original images. (b)-(g) are the heatmap visualizations of different models. M and B indicate the images with malignant tumors and benign tumors, respectively. The lesions are pointed by the white arrows.

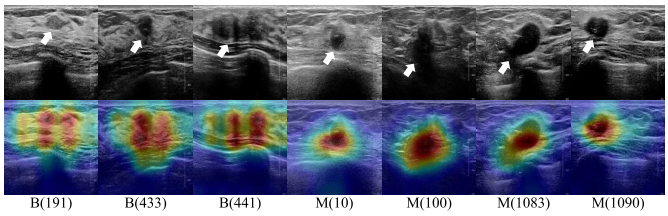


Fig. 6. More visualization of HoVer-Trans. The top row shows the original BUS images in GDPH&SYSUCC. The bottom row shows the heatmaps overlaid on the original images. M and B indicate the images with malignant tumors and benign tumors, respectively. The lesions are pointed by the white arrows.

results show that our proposed model achieves more precise diagnostic performance than two readers in our dataset. However, we also observe that the diagnostic performance of two readers is lower than the one reported in the other reference [9]. Because in this experiment, to make a fair comparison with the AI model, readers assess only one image each time. But the ultrasound examination is a dynamic process, sonographers do not just read the static images but observe the lesion from different views. Therefore, it is hard for sonographers to precisely diagnose breast lesions with only one image. It is also a limitation of this experiment. Nevertheless, our proposed model outperforms sonographers under the same condition.

Furthermore, we conduct BI-RADS subgroup analysis in the lower part in Table III. We divide the entire dataset into two subgroups by the BI-RADS scores, BI-RADS 2&3 and BI-RADS 4&5. It can be observed that the consistency between two readers is low reflected by the evaluation metrics, for example, the specificity (reader1: 0.996, reader2: 0.879) and the recall (reader1: 0.099, reader2: 0.455) in the BI-RADS 2&3 group. Such biases can be caused by different factors, such as equipment bias, cognitive bias and etc. In addition, missing clinical information may also impede the sonographers to achieve a comprehensive diagnosis. Reader1 obviously tends to classify all the lesions with BI-RADS 2 or 3 as benign tumors, which may lead to undertreatment. Under the same condition of assessing only one BUS image without any additional information, our proposed model achieves more stable diagnostic performance in both subgroups than sonographers thanks to the data-driven nature.

D. Ablation Studies

In this part, we conduct several ablation studies to evaluate the effectiveness of the anatomy-aware formulation, the association between transformer and CNN and different transformer configurations.

1) Effectiveness of Anatomical Prior Knowledge: We conduct this experiment to evaluate the effectiveness of the anatomy-aware formulation. Three variants are introduced in this experiment. 1) H branch with horizontal strip embedding and V branch with vertical strip embedding are removed, named $Model_p$. Only two main branches with patch embedding are left. 2) We remove the H branch and retain the other three branches, named $Model_{p+v}$. 3) We remove the V branch and retain the other three branches, named $Model_{p+h}$. The upper part of Table IV demonstrates the five-fold cross validation results. It can be observed that without the design of HoVer in $Model_p$, the performance of the metrics decreases by around 1%-3% except for the specificity and precision of $Model_{p+v}$. When only removing H branch or V branch, the quantitative results do not improve due to the asymmetric of the models. Fig. 7 (b)-(d) demonstrate the heatmap visualizations of three model variants and Fig. 7 (i) demonstrates our results. In Fig. 7 (b) we can observe that associating transformer with CNN can obtain visually more convincing attention maps better than transformer models only, shown in Fig. 6 (e)&(f).

TABLE IV
ABLATION STUDIES OF ANATOMY-AWARE FORMULATION, ASSOCIATION BETWEEN TRANSFORMER AND CNN AND DIFFERENT TRANSFORMER CONFIGURATIONS. (GDPH&SYSUCC)

Ablation study - Anatomy-aware formulation							
Configurations		AUC	ACC	Specificity	Precision	Recall	F1-score
$Model_p$		0.916±0.018	0.864±0.015	0.802±0.016	0.887±0.012	0.900±0.016	0.893±0.013
$Model_{p+v}$		0.919±0.022	0.880±0.018	**0.849±0.037**	**0.910±0.023**	0.898±0.018	0.904±0.015
$Model_{p+h}$		0.911±0.019	0.868±0.019	0.818±0.021	0.894±0.014	0.897±0.022	0.896±0.015
Ours		**0.924±0.016**	**0.893±0.021**	0.836±0.038	0.906±0.023	**0.926±0.019**	**0.916±0.019**
Ablation study - Convolution							
Conv Stem	Conv block	AUC	ACC	Specificity	Precision	Recall	F1-score
-	-	0.907±0.009	0.861±0.020	0.829±0.026	0.905±0.015	0.877±0.018	0.891±0.016
✓	-	0.916±0.016	0.873±0.012	**0.900±0.011**	0.827±0.026	0.826±0.013	0.826±0.010
✓	1×1	**0.926±0.015**	0.881±0.018	0.837±0.031	0.905±0.018	0.907±0.017	0.905±0.015
-	✓	0.908±0.008	0.871±0.012	0.806±0.018	0.889±0.013	0.909±0.012	0.899±0.010
✓	✓	0.924±0.016	**0.893±0.021**	0.836±0.038	**0.906±0.023**	**0.926±0.019**	**0.916±0.019**
Ablation study - Sizes of Different Embedding Ways							
p	$h\&v$	AUC	ACC	Specificity	Precision	Recall	F1-score
2	1	0.911±0.024	0.880±0.014	0.826±0.025	0.900±0.014	0.912±0.014	0.906±0.011
2	2	**0.924±0.016**	**0.893±0.021**	0.836±0.038	0.906±0.023	**0.926±0.019**	**0.916±0.019**
2	4	0.910±0.016	0.873±0.016	0.825±0.029	0.899±0.014	0.900±0.020	0.900±0.012
4	1	0.904±0.027	0.877±0.021	0.837±0.026	0.905±0.015	0.900±0.021	0.903±0.016
4	2	0.919±0.015	0.883±0.014	**0.841±0.017**	**0.908±0.010**	0.907±0.013	0.907±0.011
4	4	0.920±0.014	0.879±0.012	0.832±0.030	0.903±0.017	0.906±0.014	0.905±0.010
8	1	0.911±0.013	0.879±0.010	0.835±0.020	0.904±0.014	0.905±0.008	0.904±0.009
8	2	0.914±0.019	0.879±0.014	0.820±0.023	0.897±0.016	0.913±0.011	0.905±0.012
8	4	0.900±0.011	0.864±0.008	0.792±0.019	0.882±0.010	0.906±0.015	0.894±0.007

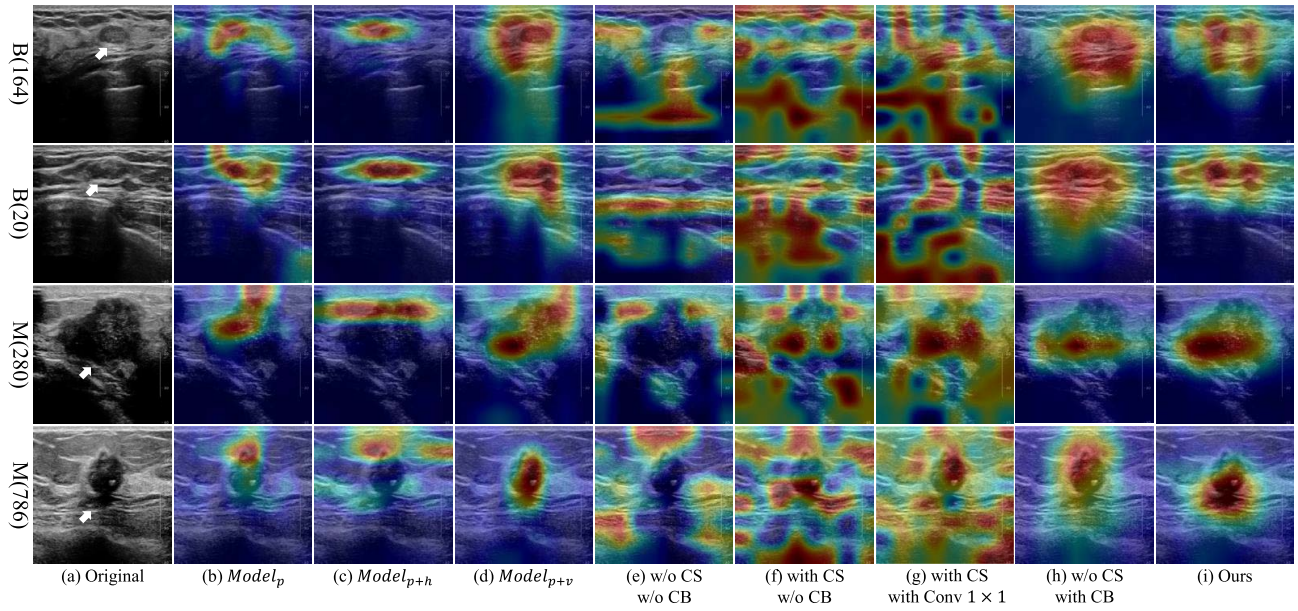


Fig. 7. The heatmaps of ablation studies. (a) shows the original images from the GDPH&SYSUCC dataset. Lesions are pointed by the white arrows. (b)-(d) are the model variants of the ablation study in the anatomy-aware formulation. (e)-(h) are the model variants of the ablation study in the convolution operations. CS and CB indicate the convolutional stem and Conv blocks, respectively. (i) shows the attention maps generated by our final model.

However, the model focuses are still imperfect due to the lack of H and V branches. When equipped with H or V branch in Fig. 7 (c)&(d), the model can focus on the anomaly regions horizontally or vertically, guided by the anatomical prior. Thanks to the complete HoVer design shown in Fig. 7 (i), our proposed model achieves the best visualization results with the most correct lesion location and attention.

2) Effectiveness of Convolution: In this experiment, we conduct an ablation study to evaluate the effectiveness of associating convolutional operations with transformer, including the early convolutional stem and the Conv blocks. We compare our model with four variants. (1) Pure transformer model without any convolutional layer. Maxpooling is applied to downsample the input image. (2) Model with conv stem only. (3) Model with conv stem and replacing the Conv block with a 1×1 convolutional layer. (4) Model with Conv block only. (5) Our final model. The quantitative results are shown in the middle part of Table IV. The attention maps are shown in Fig. 7 (e)-(i).

When without any convolutional operation in Model (1), the evaluation metrics decrease around 1%-4% with less meaningful attention maps shown in Fig. 7 (e). Introducing early

conv stem can slightly improve the classification performance, especially specificity. But the attention maps are non-local and cannot focus on the lesion. When combining the conv stem with a 1×1 convolutional layer, the quantitative performance is further improved, but the attentions maps in Fig. 7 (g) are still unsatisfactory. Because 1×1 convolutional layer only acts like channel-wise pooling for reducing the feature dimension, which is not able to aggregate the local contextual information. The above three models demonstrate the importance of introducing inductive bias. Lack of inductive bias might not greatly decrease the classification performance. But it will drastically harm the locality of the model due the self-attention mechanism in transformer. When equipped Conv blocks in Model (4), it achieves moderate quantitative performance with much better attention maps. Combining Conv blocks and the convolutional stem, the final HoVer-Trans model achieves the best classification results and the most interpretable attention maps.

3) Sizes of Different Embedding Ways: Since we introduce three embedding ways in this model to formulate the anatomical structure. In this ablation study, we further explore how the sizes of different embedding ways affect the proposed model, shown in Table IV. $p = 2$ means the visual tokens of the patching embedding are with the resolution of 2×2. $h\&v = 2$ means the tokens of the horizontal strip embedding and the vertical strip embedding are with the resolution of $2 \times width$ and $height \times 2$, respectively. In this experiment, we let $p = \{2, 4, 8\}$ and $h\&v = \{1, 2, 4\}$ and test all the combinations.

It can be observed that the classification performance of all the models is close. According to the quantitative results in Table IV, we summarize some observations on how to select these two hyper-parameters. First of all, we have to select the proper token size of the horizontal and vertical strip embeddings $h\&v$. When $h\&v = 1$, each token only contains the horizontal or vertical information with only one-pixel width or height, which is too limited. When $h\&v = 4$, the size of the token is too large, which may occupy more computational resources. Therefore, we let $h\&v = 2$ for the horizontal and vertical strip embedding. For the token size of patch embedding, we choose the value of $p = 2$ because we want to let the token of patch embedding fit the token size of horizontal and vertical strip embedding. Experimental results prove that the configuration with $p = 2$ and $h\&v = 2$ achieves the best classification performance. We apply this setting in our final model.

VI. Conclusion

In this paper, we propose a novel HoVer-Trans model, which associates the transformer with CNN, for breast cancer diagnosis in breast ultrasound images. An anatomy-aware HoVer-Trans block is designed to formulate the anatomical prior knowledge in BUS images. To achieve that, we incorporate three embedding ways, patch embedding, horizontal strip embedding and vertical strip embedding to explore spatial correlations of the inter-layer and intra-layer visual words. There are several advantages to the above technical designs. 1) The proposed model is ROI-free which does not require a pre-defined lesion ROI. Such a property greatly improves the model flexibility in clinical practice. We also believe that the whole image can deliver much more information about the peritumoral context and the spatial relationship between the lesion and different breast layers than the lesion ROIs do. 2) The proposed model can provide interpretable attention maps to support the model prediction results, which is the key that most sonographers care about when using AI algorithms to assist decision-making. 3) The proposed model achieves the best classification performance against several SOTA models in both quantitative evaluations and heatmap visualizations.

Besides, there are still several future improvements that have to be achieved. First, as we discussed in the AI vs. sonographers experiment, the breast ultrasound examination is a dynamic process. Our next plan is to aggregate the BUS images from multiple views and achieve more precise diagnostic performance, instead of just simply assessing one BUS image. Furthermore, it will be a great breakthrough if we can mimic how sonographers perform breast ultrasound examination by keeping tracking the imaging signal along with the ultrasonic probe and make a comprehensive diagnosis if the hardware is capable.

Second, due to the model complexity, the proposed model shows poor classification performance when trained by a smaller dataset, such as UDIAT. That is the reason why we also construct and release a larger dataset GDPH&SYSUCC for breast cancer diagnosis in BUS images. We are also planning to construct a way larger multi-center dataset to further explore the capacity and the generalizability of the proposed model.

Acknowledgment

Yuhao Mo is with the School of Medicine, South China University of Technology, Guangzhou, Guangdong 510006, China, and also with the Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou 510080, China (e-mail: yuhaomo3270@163.com).

Chu Han is with the Department of Radiology, Medical Research Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou 510080, China, and also with the Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China (e-mail: hanchu@gdph.org.cn).

Yu Liu, Zhenwei Shi, Bingjiang Qiu, Yanfen Cui, Lei Wu, Xipeng Pan, Zeyan Xu, Xiaomei Huang, Zhenhui Li, Zaiyi Liu, and Changhong Liang are with the Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou 510080, China, and also with the Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China (e-mail: yuyua0808@163.com; zhenwei_shi88@163.com; qiu.bingjiang@hotmail.com; wl858998458@163.com; pxp201@guet.edu.cn; zeyx0708@163.com; xmhuang1992@163.com; lizhenhui@kmmu.edu.cn; zyliu@163.com; liangchanghong@gdph.org.cn).

Min Liu is with the Department of Ultrasound, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, China (e-mail: liumin@sysucc.org.cn).

Jiatai Lin is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: linjiatai_cs@163.com).

Bingchao Zhao is with the School of Medicine, South China University of Technology, Guangzhou, Guangdong 510006, China, and also with the Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou 510080, China (e-mail: zbcajj@163.com).

Chunwang Huang is with the Department of Ultrasound, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou 510080, China (e-mail: huangchunwang@126.com).

Ying Wang is with the Department of Medical Ultrasonics, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China (e-mail: liuivy527@163.com).

REFERENCES

[1] H. Sung et al., "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] N. J. Massat, A. Dibden, D. Parmar, J. Cuzick, P. D. Sasieni, and S. W. Duffy, "Impact of screening on breast cancer mortality: The U.K. program 20 years on," *Cancer Epidemiol., Biomarkers Prevention*, vol. 25, no. 3, pp. 455–462, Mar. 2016.

[3] T. B. Bevers et al., "Breast cancer screening and diagnosis, version 3.2018, NCCN clinical practice guidelines in oncology," *J. Nat. Comprehensive Cancer Netw.*, vol. 16, no. 11, pp. 1362–1389, 2018.

[4] X. Qian et al., "Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 522–532, Jun. 2021.

[5] Y. Wang et al., "3D inception U-Net with asymmetric loss for cancer detection in automated breast ultrasound," *Med. Phys.*, vol. 47, no. 11, pp. 5582–5591, Nov. 2020.

[6] Y. Wang et al., "Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 866–876, Apr. 2020.

[7] Y. Zhou et al., "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101918.

[8] X. Qian et al., "A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network," *Eur. Radiol.*, vol. 30, no. 5, pp. 3023–3033, May 2020.

[9] Y. Shen et al., "Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams," *Nature Commun.*, vol. 12, no. 1, pp. 1–13, Sep. 2021.

[10] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[11] M. H. Yap et al., "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1218–1226, Jul. 2017.

[12] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2713–2724, Sep. 2020.

[13] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2017.

[14] S. Graham et al., "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101563.

[15] A. E. Ilesanmi, U. Chaumrattanakul, and S. S. Makhanov, "Methods for the segmentation and classification of breast ultrasound images: A review," *J. Ultrasound*, vol. 24, no. 4, pp. 367–382, Dec. 2021.

[16] Y. Liu et al., "A survey of visual transformers," 2021, *arXiv:2111.06091*.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2016.

[18] Z. Ning, C. Tu, Q. Xiao, J. Luo, and Y. Zhang, "Multi-scale gradational-order fusion framework for breast lesions classification using ultrasound images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2020, pp. 171–180.

[19] X. Zheng et al., "Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer," *Nature Commun.*, vol. 11, no. 1, pp. 1–9, Mar. 2020.

[20] X. Guo et al., "Deep learning radiomics of ultrasonography: Identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer," *EBioMedicine*, vol. 60, Oct. 2020, Art. no. 103018.

[21] M. Jiang et al., "Deep learning with convolutional neural network in the assessment of breast cancer molecular subtypes based on U.S. images: A multicenter retrospective study," *Eur. Radiol.*, vol. 31, no. 6, pp. 3673–3682, Jun. 2021.

[22] M. Byra et al., "Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion," *Med. Phys.*, vol. 46, no. 2, pp. 746–755, Feb. 2019.

[23] W. K. Moon, Y.-W. Lee, H.-H. Ke, S. H. Lee, C.-S. Huang, and R.-F. Chang, "Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105361.

[24] Z. Zhuang, Y. Kang, A. N. Joseph Raj, Y. Yuan, W. Ding, and S. Qiu, "Breast ultrasound lesion classification based on image decomposition and transfer learning," *Med. Phys.*, vol. 47, no. 12, pp. 6257–6269, Dec. 2020.

[25] W. Cui et al., "FMRNet: A fused network of multiple tumoral regions for breast tumor classification with ultrasound images," *Med. Phys.*, vol. 49, no. 1, pp. 144–157, Jan. 2022.

[26] B. Zhang, A. Vakanski, and M. Xian, "Bi-RADS-Net: An explainable multitask learning approach for cancer diagnosis in breast ultrasound images," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2021, pp. 1–6.

[27] E. Zhang, S. Seiler, M. Chen, W. Lu, and X. Gu, "BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis," *Phys. Med. Biol.*, vol. 65, no. 12, Jun. 2020, Art. no. 125005.

[28] X. Di, S. Zhong, and Y. Zhang, "Saliency map-guided hierarchical dense feature aggregation framework for breast lesion classification using ultrasound image," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106612.

[29] X. Qi et al., "Automated diagnosis of breast ultrasonography images using deep neural networks," *Med. Image Anal.*, vol. 52, pp. 185–198, Feb. 2019.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[31] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, *arXiv:2106.04554*.

[32] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[33] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[34] F. Shamshad et al., "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[35] X. Gao, Y. Qian, and A. Gao, "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models," 2021, *arXiv:2107.01682*.

[36] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.

[37] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[38] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollar, and R. Girshick, "Early convolutions help transformers see better," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Red Hook, NY, USA: Curran Associates, 2021, pp. 30392–30400. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–14.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[42] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[43] J. Xing et al., "Using BI-RADS stratifications as auxiliary information for breast masses classification in ultrasound images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2058–2070, Jun. 2021.