IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 July 2024, accepted 11 August 2024, date of publication 19 August 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445954

RESEARCH ARTICLE

ST-Double-Net: A Two-Stage Breast Tumor Classification Model Based on Swin Transformer and Weakly Supervised Target Localization

SHENGNAN HAO[1], YIHAN JIA[1], JIANUO LIU[1], ZHIWU WANG[2,3], CHUNLING LIU[2,3], ZHANLIN JI[4,5], (Member, IEEE), AND IVAN GANCHEV[5,6,7], (Senior Member, IEEE)

[1]Department of Artificial Intelligence, North China University of Science and Technology, Tangshan 063009, China
[2]Hebei Key Laboratory of Molecular Oncology, Tangshan, Hebei 063001, China
[3]Tangshan People's Hospital, Tangshan, Hebei 063001, China
[4]College of Mathematics and Computer Science, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China
[5]Telecommunications Research Centre (TRC), University of Limerick, Limerick, V94 T9PX Ireland
[6]Department of Computer Systems, University of Plovdiv "Paisii Hilendarski," 4000 Plovdiv, Bulgaria
[7]Institute of Mathematics and Informatics—Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria

Corresponding authors: Chunling Liu (lcl.1983.hi@163.com), Zhanlin Ji (zhanlin.ji@gmail.com), and Ivan Ganchev (ivan.ganchev@ul.ie)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0135700; in part by the Bulgarian National Science Fund (BNSF) under Grant КП-06-ИП-КИТАЙ/1 (KP-06-IP-CHINA/1), and in part by the Telecommunications Research Centre (TRC) of University of Limerick, Ireland.

ABSTRACT Breast cancer is the second deadliest cancer (after lung cancer) globally among women, with high incidence and mortality rates. Its early diagnosis is pivotal for improving the cure rate. With the continuous development and maturity of deep learning technologies, traditional classification models have been widely applied for automated classification of pathological images. However, several challenges still persist. For instance, traditional classification models typically perform well in processing images with clear distinctions between target objects and backgrounds, but struggle to accurately classify pathological images due to the lack of clear distinctions between tumor lesion areas and background areas. In the light of this, we propose a two-stage breast tumor pathological classification model based on weakly supervised target localization, named ST-Double-Net. In the proposed model, precise lesion localization and classification are achieved in two stages. In the first stage, a set of *global* feature maps is obtained by utilizing the Swin Transformer. These feature maps are then input into a newly designed heatmap cropping (HMC) module, which forces the model to focus on discriminative features of lesion areas through heatmap-guided cropping, without requiring bounding boxes or relevant annotation information. This gradual refinement of target localization facilitates the extraction of useful global features, from coarse to fine. The images with discriminative features generated in the first stage serve as inputs for the second stage, where another Swin Transformer extracts *local* features from the magnified lesion region images. Finally, the global and local features extracted in the first and second stage, respectively, are fused to emphasize subtle differences in the images, thereby enhancing the model's classification ability. The proposed ST-Double-Net model is evaluated on the BreaKHis and BACH public datasets, demonstrating superior performance compared to state-of-the-art models.

INDEX TERMS Breast tumor classification, Swin Transformer, weakly supervised target localization, heatmap cropping.

I. INTRODUCTION

Breast cancer is considered one of the most prevalent cancer types globally, with incidence counts steadily rising in recent years. Clinical studies underscore the close correlation between early detection and cure rates, emphasizing the critical importance of early diagnosis. Various methods, based on X-ray imaging [1], magnetic resonance imaging [2], ultrasound imaging [3], and biopsies [4], are commonly employed for early breast cancer diagnosis. Among these,

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

biopsy stands as the only method capable of accurately localizing lesions and completing the tumor subtype classification. With the continuous advancement of deep learning technologies, an increasing number of convolutional neural networks (CNNs), such as ResNet [5] and DenseNet [6], have been proposed. These networks have propelled research into breast cancer pathological image classification tasks. However, in traditional image classification networks, regardless of the proportion of important discriminative regions within an image, only a uniform feature extraction is applied to the entire image. Unlike images with clear distinctions between target objects and backgrounds, in breast cancer histopathological images, target objects and backgrounds are intertwined without clear boundaries, making it challenging to localize lesion areas [7]. This characteristic of pathological images poses difficulties for traditional classification networks in accurately extracting features of lesion areas, thereby increasing the complexity of image classification.

To enhance the extraction of crucial features from lesion areas, breast cancer pathological images are typically classified using a localization-identification approach. Analogous to how humans differentiate similar objects, this approach often involves initially identifying regions of interest through rapid scanning, followed by feature extraction and meticulous comparison for classification. However, current localization-classification methods face two main challenges related to the accurate localization of key regions and effective feature extraction from these regions. Huang et al. [8] proposed a Part-Stacked CNN (PS-CNN) for fine-grained visual categorization, based on manually labeled strong component annotations, utilizing a fully convolutional network for component localization and a dual-stream classification network for encoding object- and component features. This strong supervised approach, aided by annotated bounding boxes, detects key regions, thereby mitigating the influence of other noisy features and reducing the difficulty of region localization. Nevertheless, such strong supervised methods typically require additional annotation information, such as more bounding boxes, to enable the network to learn the positional information of target regions. This process consumes more human resources for image annotation and, in addition, the manually annotated positions may not necessarily represent the most discriminative regions, thus leading to over-reliance on the annotators' cognitive level, which significantly impacts classification results [9].

To address these challenges, this paper proposes a two-stage breast tumor classification model, named ST-Double-Net, utilizing a newly designed heatmap cropping (HMC) module with weakly supervised target localization [10], whereby target regions are located without the need for bounding boxes or other annotation information, using only weakly supervised methods. The HMC module utilizes the Gradient-weighted Class Activation Mapping (Grad-CAM) [11] to analyze the network's focus areas for a particular class, solely based on image class labels, and to visualize the areas of focus during a Swim Transformer training, enabling the localization of key regions. Subsequently, image cropping based on the identified key regions amplifies the discriminative features of subtle regions. The weakly supervised target localization approach utilized by the HMC module enables the rough localization of target objects at minimal cost without requiring additional bounding box information [12]. By employing a two-stage classification method combined with the HMC module, the proposed ST-Double-Net model can progressively refine the target localization in order to extract useful features. The primary goal of the first stage is to refine the original images by using the Swin Transformer for *global* feature extraction and focusing on subtle discriminative features through the HMC module. The refined lesion area images output by the first stage are then input into the second stage for *local* feature extraction. Finally, the fusion of multi-scale features, extracted by the two stages, captures subtle differences between images, thus reducing the difficulty of classification.

The proposed ST-Double-Net model achieves automated identification and classification of lesion areas, serving as an auxiliary tool for pathologists to enhance the efficiency of their work. Traditional pathological analysis and diagnosis require substantial human resources and manual annotations. Each pathological diagnostic result must be meticulously observed by pathologists, which undoubtedly increases their workload. Furthermore, diagnostic results may vary among different pathologists due to differences in their experience and skills. Therefore, the utilization of the proposed two-stage classification model can provide a valuable help to pathologists, allowing them to combine the automated classification results with their professional knowledge for making more accurate judgments. This approach also helps reduce subjective biases in diagnoses. Particularly in complex and challenging cases, automated classification results can offer more valuable reference information. Therefore, the introduction of ST-Double-Net in practice can contribute to more accurate diagnosis of early-stage breast cancer pathologists, thereby improving cure rates, which holds significant clinical implications.

With the advent of computer-aided treatment, integrating ST-Double-Net into existing diagnostic workflows to enhance the diagnostic efficiency is essential. Initially, a sufficient number of pathological images should be collected in clinical environments. After undergoing standardized preprocessing, these images can be used for better training the model. After that, the mature diagnostic model can be deployed into the existing diagnostic systems ensuring pathologists can easily access and interpret the model's results. During clinical diagnosis, pathologists could upload processed pathological images to the diagnostic systems, which then can provide classification diagnostic reports based on the input images. Pathologists can review and verify these results by applying their professional knowledge and experience to make final diagnostic decisions. In summary, the

proposed ST-Double-Net model can improve the accuracy and efficiency of early diagnosis of breast cancer, reduce associated costs, and provide strong support for pathologists' diagnostic decisions.

In summary, this paper makes the following contributions:

1) Proposing a newly designed heatmap cropping (HMC) module to address the issue of traditional networks overlooking subtle features. By using a weakly supervised approach, lesion areas are localized without requiring bounding box annotations.

2) Proposing a two-stage classification method combined with the newly designed HMC module in a phased manner to extract useful features, from coarse to fine. By combining coarse-grained features with fine-grained features, inter-class differences are captured effectively, enabling the extraction of lesion area features and improving the classification performance.

3) Testing the proposed ST-Double-Net model on two publicly available and widely used datasets, namely the Breast Cancer Histopathological Image Classification (BreaKHis) [13] and BreAst Cancer Histology images (BACH) [14], in a comparing manner to existing advanced models, demonstrating its superiority.

II. RELATED WORK

Cancer is the main killer, threatening human health, among which breast cancer is one of the most common types. The incidence of breast cancer is very high in women, and its early diagnosis is crucial to improve the cure rate [15]. Compared with the simple classification of benign and malignant breast cancer, the accurate identification of breast tumor subtypes can help doctors choose better treatment plans. With the development of deep learning technologies, a variety of traditional CNNs have been widely used in the classification of breast cancer pathological images.

Vang et al. [16] proposed a deep learning framework to solve the problem of multi-class classification of breast cancer histopathological images. These authors use CNNs to learn and extract features from image samples. Their proposed framework uses a breast cancer histopathological image dataset for training and achieves good performance in multi-class classification tasks. Ting et al. [17] proposed a CNN model that addresses some of the issues in traditional CNN models as to better process pathological images. The effectiveness of the proposed model is verified by experiments on a breast cancer histopathological image dataset. Ragab et al. [18] introduced a framework for classification of pathological images of breast cancer using multiple Deep CNNs (DCNNs). Their proposed framework is experimentally evaluated, using a breast cancer pathological image dataset, in comparison with existing models. The obtained results demonstrate that the method of integrating multiple DCNNs achieved good performance in pathological image classification tasks, showing good application prospects. Using a public breast cancer pathological image dataset, Liew et al. [19] investigated the performance of an XGBoost-based breast tumor classification algorithm in the classification task of breast cancer pathological images, evaluated its classification effect, and compared it with other commonly used classification algorithms, fully verifying its robustness. Jiang et al. [20] proposed the Breast TransFG Plus model, which is based on the Transformer architecture and addresses the needs of fine-grained classification tasks. The model uses a multi-head self-attention mechanism to capture subtle features in images and enhance its ability to recognize different tissue structures in pathological images. Extensive experiments, conducted on multiple public breast cancer pathology image datasets, demonstrated that the Breast TransFG Plus model is superior to traditional models in terms of classification accuracy and robustness.

Although Transformers have the advantage of capturing global features, when dealing with larger images each pixel needs to be processed with other pixels in the image, resulting in massive computational overhead. Recognizing this challenge, Liu et al. [21] proposed the Swin Transformer. The core innovation of Swin Transformer lies in its design of window shifting, which avoids computation of global sub-attentions while reinforcing inter-window connections. The Swin Transformer architecture is illustrated in Figure 1, comprising four pivotal stages, each progressively reducing the resolution of input feature maps and enlarging receptive fields layer by layer, akin to CNNs. The computation process of consecutive Swin Transformer blocks can be represented, as shown in [21], as follows:

$$\hat{z}^{l} = WMSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \tag{1}$$

$$z^{l} = MLP\left(LN\left(\hat{z}^{l}\right)\right) + \hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1} = SWMSA\left(LN\left(z^{l}\right)\right) + z^{l} \tag{3}$$

$$z^{l+1} = MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{4}$$

where $\hat{z}^{l}$ and z^{l} denote the output features of the (Shifted) Window based Multi-head Self-Attention, (S)WMSA, module and the Multi-Layer Perceptron (MLP) module for block l, and LN denotes a Layer Normalization (LN). WMSA is an attention mechanism introduced in the Swin Transformer to improve its computing efficiency and locality by applying self-attention within a fixed-size window. MLP is a classic neural network component that usually contains one or more fully connected (FC) layers and nonlinear activation functions. LN is a method used to standardize the output of neural network layers, aiming to speed up the model training process and improve the performance. SWMSA is an improved attention mechanism introduced in the Swin Transformer to enhance the modeling ability of attention within the window through the sliding window method.

From the above research results, it can be found that many models have been proposed and great progress has been made in the task of breast cancer pathological image

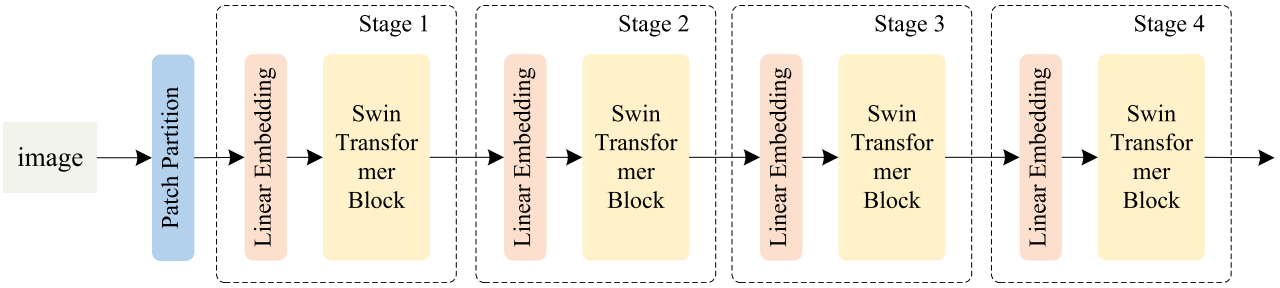


FIGURE 1. The Swin Transformer, utilized by the proposed ST-Double-Net model.

classification. However, there are still many issues that need to be addressed. For example, when using histopathological images for accurate diagnosis, it is necessary to make judgments based on the morphology of tissue cells in the lesion area [22]. However, pathological images are different from ordinary images. The difference between the target object and background in ordinary images is very obvious, with clear boundary lines, making it easy to classify. In pathological images, the target object and the background are often mixed together, and the morphological differences of tissue cells are minor. Therefore, it is very important to obtain the accurate positioning of the lesion area and extract the corresponding key features [23].

By using the combination of visual and linguistic information for fine-grained recognition, in 2020 Song et al. [24] proposed a Progressive Masked Attention (PMA) model that can be trained end-to-end, requiring only original images and text descriptions, whose accuracy was proved by comprehensive experiments conducted on fine-grained benchmark datasets. Branson et al. [25] trained a network for object pose estimation by labeling points and anchor boxes, which made it easier to learn the characteristics of birds without the interference of pose changes, and achieved good classification performance on multiple datasets.

While leveraging additional information, such as bounding boxes, which can effectively improve the classification performance, such strong supervised classification methods severely consume human resources. To meet the needs of practical applications, weakly supervised methods are often used to achieve localization and classification of key regions. Ke et al. [26] proposed an end-to-end two-level attention activation model (TL-AAM), applying object attention activation modules to complementarily locate object regions. Using a multi-scale pyramid attention localization module, the feature channel with the maximum response value is selected to locate local feature regions. A substantial number of experiments confirmed the usability of this model. In 2020, Zhang et al. [27] introduced a method for fine-grained classification using weakly supervised part detection networks, designing a novel network architecture that can simultaneously perform object detection and fine-grained classification tasks without requiring part-level annotations, and achieved promising results on different datasets. Patil et al. [28] proposed an attention-based multi-instance learning model for better localization of malignant regions in breast tissue pathology images. This model treats the image classification problem as a weakly supervised learning problem, considering only image-level labels and not the exact locations of the cancerous areas. It divides each image into multiple small patches, extracts features from each small patch, and utilizes an attention mechanism to weight the importance of each small patch in classification. Fan et al. [29] introduced a model for microscopic fine-grained instance classification, implemented through a deep attention mechanism. This model first uses a lightweight gated attention mechanism to detect multiple discriminative regions, and then combines global structure and local instance features for final image-level classification. Experimental results show that it is effective for breast tumor classification. Generally, the weakly supervised target localization reduces the dependence on annotated data, improves the data utilization efficiency and generalization ability of models, thus providing a more economical and practical solution for practical applications. Therefore, it is very promising to apply it to the classification task of breast cancer pathological images.

Through the study of various models, it can be observed that, compared to traditional supervised learning methods, achieving localization of target regions in a weakly supervised manner exhibits better generalization ability and robustness [30]. In addition, it helps address the problem of scarce part-level annotation data in various classification tasks. Based on this, the current paper proposes a two-stage classification model based on weakly supervised target localization, which achieves accurate localization and amplification of lesion areas without requiring additional bounding box information, thus greatly enhancing the classification performance.

III. PROPOSED MODEL

A. OVERALL STRUCTURE

Due to the high similarity between target objects and backgrounds in histopathological images, discernment by the human eye is challenging, making the classification of breast cancer pathology images a formidable task. Most existing models adopt a traditional end-to-end training approach, where all features in the images are indiscriminately

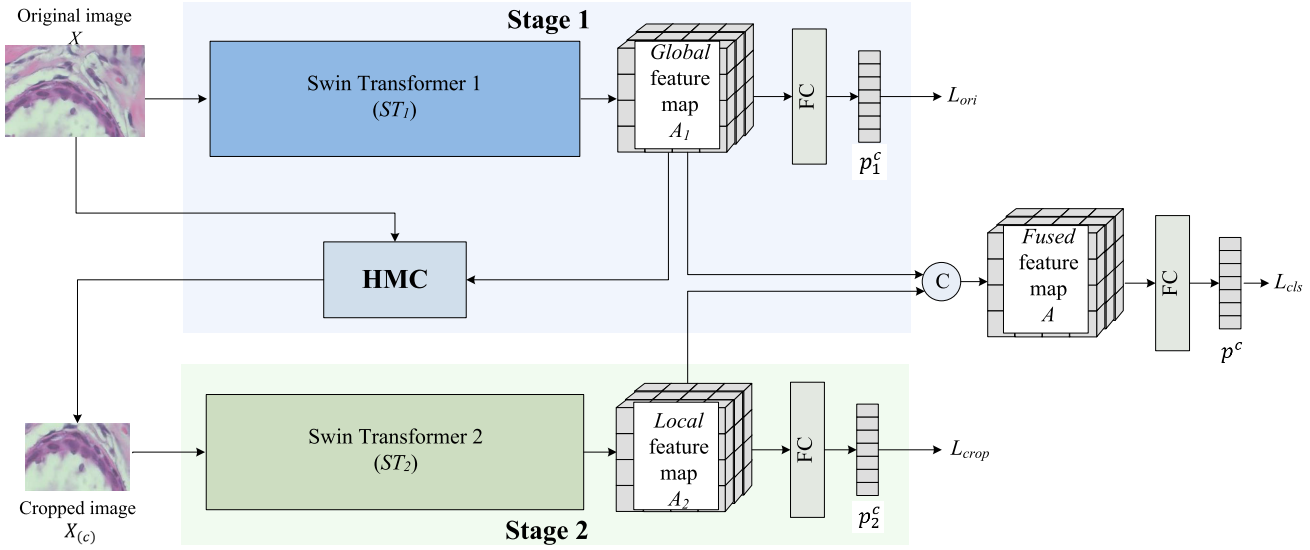


FIGURE 2. The proposed ST-Double-Net model.

extracted. However, the unclear boundaries in histopathological images often lead to difficulties in locating target lesion areas, thus significantly affecting classification outcomes.

To address these issues, a two-stage classification model, based on the Swin Transformer, is proposed in this paper, utilizing the locate-first-then-classify approach. The proposed ST-Double-Net model is depicted in Figure 2. In the first stage, it locates the tumor lesion areas by employing a Swin Transformer to extract *global* feature information and obtain a corresponding set of feature maps. Then, by employing a weakly supervision, it achieves precise localization and magnification of suspicious regions without a need of bounding boxes, by gradually refining target localization and extracting discriminative *global* features, from coarse to fine. In the second stage, the model utilizes another Swin Transformer to extract *local* features of the lesion areas. Finally, the extracted *global* features and *local* features are fused. The fused features contain more semantic information, which allows to greatly improve the classification performance of the model.

B. STAGE 1

In Figure 2, an original image X is input into the first Swin Transformer (ST_1), utilized to extract the *global* features.

Via a forward propagation, the predicted value p_1^c for the image's class c is obtained as follows:

$$p_1^c = ST_1\left(W_t * X\right) \tag{5}$$

where W_t denotes the weight matrix. Subsequently, in the HMC module (described in detail in the second subsection below), the gradient of the target class c with respect to the last convolutional layer of ST_1 is computed via backpropagation. The gradient is then multiplied by the output feature map of the last convolutional layer to obtain weighted gradients for each feature map channel, reflecting the importance of each channel for the target class. These weighted gradients are aggregated to generate a heatmap, where each pixel indicates the activation level for the target class. Higher values indicate greater model attention to the class at that position. Based on the generated heatmap by means of Grad-CAM [11], the position and size of the regions of interest can be determined using a thresholding method. The regions with higher pixels in the heatmap are selected as the cropping regions, and the corresponding cropping operation is applied to the original image to extract the areas of interest. The cropped image $X_{(c)}$ is obtained as follows:

$$X_{(c)} = cam\left(X\right) \Theta X \tag{6}$$

where $cam\left(\cdot\right)$ denotes the process of heatmap generation, while Θ signifies the region cropping operation.

1) LOSS FUNCTION

The first stage adopts the Cross Entropy loss [31], calculated as follows:

$$L_{ori} = -\frac{1}{N}\sum_i \sum_{c=1}^{M} y_{ic} \log\left(p_i^c\right) \tag{7}$$

where M denotes the total number of classes, N denotes the total number of samples, y_{ic} denotes the actual class of sample i, and p_i^c denotes the predicted probability of sample i belonging to class c.

2) HMC MODULE

To address the challenge of distinguishing between the target object and background in pathological images, a method of classifying breast cancer pathological images, using a locate-first-then-classify approach, is employed. Existing locate-first-then-classify methods commonly rely on precise

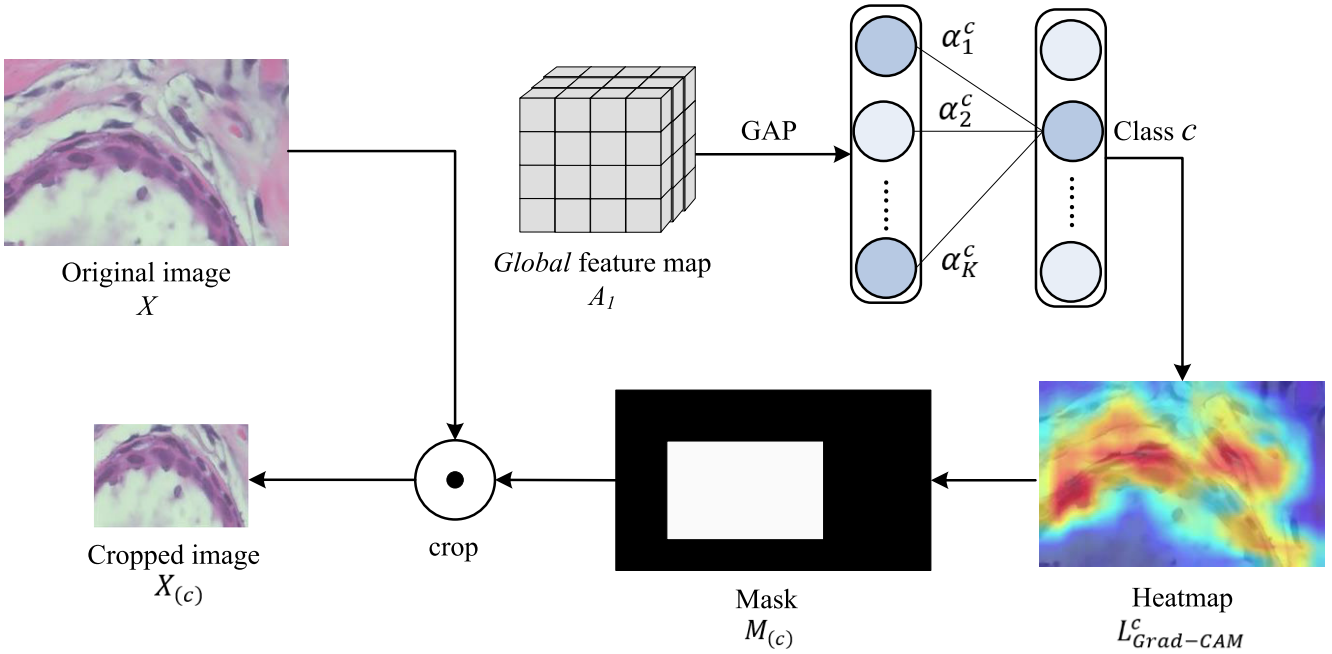


FIGURE 3. The newly designed HMC module, utilized by the proposed ST-Double-Net model.

part localization using annotated boxes to facilitate easier classification. However, this approach heavily relies on accurately annotated image parts, which requires significant human effort and impedes the practical application of pathological image classification. To tackle this issue, we propose here the use of a newly designed HMC module based on weakly supervised target localization.

As illustrated in Figure 3, the HMC module achieves the localization and amplification of crucial features using only image-level labels, without the need for annotated boxes. It utilizes Grad-CAM [11] a technique for visualizing the most attended regions in deep learning models. It generates a heatmap by backpropagating the class gradients into the feature maps of a CNN, and then weighting and summing these gradients to show the model's focus areas in the input image. Subsequently, based on the obtained heatmap, cropping operations are performed on the original image to precisely localize and amplify discriminative features in the lesion area.

For a given image $X \in R^{c\times h\times w}$, the output on the last convolutional layer A of ST_1 is fed through a fully connected (FC) layer to obtain (by forward propagation) the predicted value y^c for the image's class c. By backpropagation, according to y^c, the gradient information of the last feature layer A is obtained, and based on this, the importance of each channel in feature map A_1, denoted as α_k^c, is calculated using the following formula from [11]:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{8}$$

where A_{ij}^k denotes the value at coordinates (i, j) on the k-th channel of feature layer A, $\frac{\partial y^c}{\partial A_{ij}^k}$ signifies the gradient information of class c backpropagated onto feature layer A, and Z denotes the product of width w and height h. During this process, the computed gradients are globally average-pooled (GAP) over the width i and height j dimensions, resulting in importance weights α_k^c.

After obtaining the weights of y^c for each feature channel in A, each feature map is multiplied by its corresponding weight and summed up. Then, the ReLU activation function is applied to obtain the output Grad-CAM heatmap $L^c_{Grad-CAM}$ for class c, as per [11], as:

$$L^c_{Grad-CAM} = ReLU\left(\sum\nolimits_k \alpha_k^c A^k\right) \tag{9}$$

where A^k denotes the weight matrix on the k-th channel of feature layer A. The ReLU activation function is utilized to ensure that the output is greater than 0, suppressing uninteresting weight portions.

Then, having obtained the heatmap $L^c_{Grad-CAM}$ for class c, which accomplishes the localization of key features, the original image is cropped based on this localization.

Given the original image X and its heatmap $L^c_{Grad-CAM}$ for class c, the values in $L^c_{Grad-CAM}$ are compared with a threshold τ to determine which regions contain critical features. Let $L^c_{Grad-CAM}(i, j)$ denote the pixel value at index (i, j) in $L^c_{Grad-CAM}$. Then, the cropping process can be expressed as follows:

$$M_{(c)}(i, j) = \begin{cases} 1, if \;\; L^c_{Grad-CAM}(i, j) > \tau \\ 0, otherwise \end{cases} \tag{10}$$

where $M_{(c)}(i,j)$ denotes a binary mask indicating whether the region at position (i,j) contains selectively significant features. The original image X is cropped using the binary mask $M_{(c)}$ to amplify the selective regions as follows:

$$X_{(c)}(i,j) = X(i,j) \times M_{(c)}(i,j) \tag{11}$$

where $X_{(c)}(i,j)$ denotes the region at position (i,j) of the cropped image and $X(i,j)$ denotes the pixel value at position (i,j) in the original image. Following this cropping process, the amplification of critical regions is achieved, enabling the model to focus on subtle features that are difficult to discern, thereby enhancing the model classification performance.

C. STAGE 2

The cropped image $X_{(c)}$ obtained in the first stage is used as the input to a second Swin Transformer (ST_2) performing *local* feature extraction, whereby the predicted value p_2^c for the image's class c is obtained as follows:

$$p_2^c = ST_2\left(W_t * X_{(c)}\right) \tag{12}$$

where W_t denotes the weight matrix.

The loss function adopted in the second stage is the Focal Loss [32], calculated as follows:

$$L_{crop} = -(1-p_t)^{\gamma} \log(p_t) \tag{13}$$

where p_t denotes the model's predicted probability for the sample, and γ is an adjustable hyperparameter. This procedure tackles the issue of traditional networks treating all features equally during extraction.

After completing both stages, the extracted feature maps A_1 and A_2 are fused along the channel dimension to obtain the *fused* feature map A (c.f., Figure 2), as follows:

$$A = A_1 \oplus A_2 \tag{14}$$

where $\oplus$ indicates that the two feature maps are subjected to a feature fusion operation on the channel. This process realizes the fusion of *local* and *global* features in a pathological image, and then inputs the fused features into the classification head to obtain the final prediction result p^c of the original image X, as follows:

$$p^c = Softmax\left(W_c \cdot \sigma\left(W \cdot A + b\right) + b_c\right) \tag{15}$$

where W denotes the weight matrix of the fully connected (FC) layer, b denotes the bias vector, σ denotes the ReLU activation function, W_c denotes the weight matrix of the classification layer, and b_c denotes the bias vector of the classification layer.

The loss function used in this process is the cross-entropy loss function [31] L_cls.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

The first dataset, used in the experiments, was the publicly available and widely used [33], [34] BreaKHis dataset [13], comprising a total of 7,909 microscopic images (in *.png* format with 700 × 460 pixels, 3-channel RGB, 8-bit depth per channel) of breast tumor tissue collected anonymously from 82 patients by means of surgical open biopsy (SOB) in 2014. The images were obtained at four magnification levels (40 ×, 100 ×, 200 ×, and 400 ×) and contain two tumor classes, benign and malignant, with 2480 benign samples and 5429 malignant samples (Table 1). Among these, the benign class is subdivided into four subclasses – adenosis (A), fibroadenoma (F), phyllodes adenoma (PT), and tubular adenoma (TA), whereas the malignant class contains four subclasses – ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). Sample images of these are depicted in Figure 4. In the experiments, this dataset was randomly divided into a training set and a test set according to the ratio of 8:2.

TABLE 1. Classification of images in BreaKHis dataset.

Magnification	Benign-class images				Malignant-class images				Total images
	A	F	PT	TA	DC	LC	MC	PC	
40 ×	114	253	109	149	864	156	205	145	**1995**
100 ×	113	260	121	150	903	170	222	142	**2081**
200 ×	111	264	108	140	896	163	196	135	**2013**
400 ×	106	237	115	130	788	137	169	138	**1820**
Total images	**444**	**1014**	**453**	**569**	**3451**	**626**	**792**	**560**	**7909**

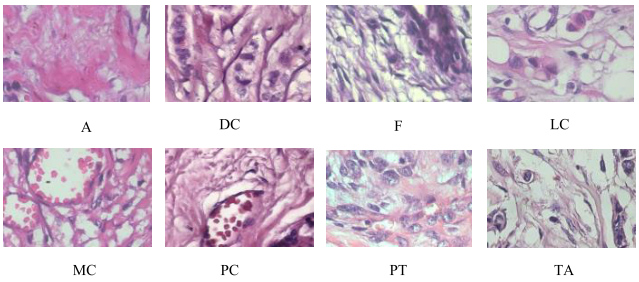


FIGURE 4. Sample BreaKHis images of eight subclasses of breast tumors.

To demonstrate the generalization ability of the proposed model, a second publicly available and widely used [33], [34] dataset was used, namely the BACH dataset [14], containing 400 high-resolution (2048 × 1536 pixels) breast histology microscopy images divided into four classes: normal, benign, in-situ carcinoma, and invasive carcinoma (Table 2). Sample images of these are shown in Figure 5. In the experiments, this dataset was randomly divided into a training set and a test set in a ratio of 8:2.

Data augmentation techniques were employed to address the problem of imbalanced sample distribution in the utilized public datasets, ensuring a balanced sample distribution across all classes and enhancing the model generalization ability.

TABLE 2. Classification of images in BACH dataset.

Class	Images
Normal	100
Benign	100
In-situ carcinoma	100
Invasive carcinoma	100
Total images	**400**

B. EXPERIMENTAL SETUP

PyTorch version 2.0.0, Python version 3.8, and CUDA version 12.1 were used for conducting the experiments. All experiments were performed on a host computer with NVIDIA GeForce RTX 3060 and Intel(R) Xeon(R) CPU E5-2686 v4 CPU @ 3.0GHz with 12G video memory.

In the classification task, the initial learning rate was set to 0.0005, the number of epochs was set to 100, and the batch size was set to 16. The Stochastic Gradient Descent (SGD) optimizer [35] was used to optimize the model. The momentum was set to 0.9 and the weight decay was 0.0001.

C. EVALUATION METRICS

For model performance evaluation, in the experiments we used four popular metrics, namely accuracy, precision, recall, and F1-score.

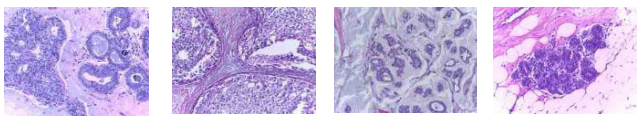


FIGURE 5. Sample BACH images of four classes of breast tumors.

Accuracy refers to the ratio of the number of samples correctly classified by a model to the total number of samples, as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{16}$$

where True Positives (TP) denotes the number of positive classes correctly predicted by a model, False Negatives (FN) denotes the number of negative classes incorrectly predicted by a model, False Positives (FP) denotes the number of positive classes incorrectly predicted by a model, and True Negatives (TN) denotes the number of negative classes correctly predicted by a model.

Representing the proportion of the samples correctly predicted as positive classes by a model to the total number of samples predicted as positive class, precision is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

Recall represents the proportion of samples that are correctly classified as positive among the number of samples that are truly positive, as follows:

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

F1-score is the harmonic mean of precision and recall, calculated as follows:

$$F1 = \frac{2.precision.recall}{precision + recall} \tag{19}$$

D. CHOICE OF FEATURE EXTRACTION NETWORK

As mentioned before, we chose the Swin Transformer [21] as the feature extraction network for the proposed two-stage classification model, based on the experimental results presented in Tables 3 and 4, comparing it to four other networks, namely ResNet-18 [5], ConvNeXt-tiny [36], EfficientNet-v2 [37], and MobileNet-v2 [38], on the BreaKHis and BACH datasets, respectively. From these tables, it is evident that the Swin Transformer significantly surpasses all other networks, according to all evaluation metrics (except for only two cases). Therefore, in the subsequent construction of the proposed two-stage classification model, the Swin Transformer was selected as the feature extraction network in its both stages.

TABLE 3. Multi-class classification performance of different networks at various magnification levels on BreaKHis dataset.

Model	Magnification	Evaluation metrics (%)			
		Overall accuracy	Average precision	Average recall	Average F1-score
ResNet-18	40 ×	94.18	93.75	95.25	94.13
	100 ×	91.55	89.75	90.13	90.00
	200 ×	89.25	89.00	84.88	86.38
	400 ×	88.64	89.25	83.50	85.63
ConvNeXt-tiny	40 ×	93.16	92.13	94.25	93.00
	100 ×	92.27	91.63	**92.50**	91.88
	200 ×	91.00	**91.13**	87.00	88.50
	400 ×	90.86	90.00	86.63	88.13
EfficientNet-v2	40 ×	87.09	86.46	87.13	85.75
	100 ×	84.54	80.88	80.13	80.25
	200 ×	87.50	85.63	80.75	83.25
	400 ×	85.32	84.63	77.75	80.00
MobileNet-v2	40 ×	80.51	80.38	78.13	78.63
	100 ×	79.95	80.75	72.63	74.75
	200 ×	78.75	80.00	68.75	71.38
	400 ×	78.40	78.13	69.63	72.25
Swin Transformer	40 ×	**95.95**	**95.99**	**95.75**	**95.78**
	100 ×	**93.72**	**93.65**	91.69	**92.46**
	200 ×	**93.00**	90.41	**91.33**	**90.82**
	400 ×	**92.52**	**94.50**	**89.75**	**91.38**

TABLE 4. Multi-class classification performance of different networks on bach dataset.

Model	Evaluation metrics (%)			
	Overall accuracy	Average precision	Average recall	Average F1-score
ResNet-18	87.50	88.08	87.50	87.00
ConvNeXt-tiny	91.25	91.29	91.25	91.21
EfficientNet-v2	90.00	90.57	90.00	90.14
MobileNet-v2	90.00	90.20	90.00	89.91
Swin Transformer	**92.50**	**92.55**	**92.50**	**92.50**

E. MULTI-CLASS CLASSIFICATION

1) ON BreaKHis DATASET

The confusion matrices of the proposed ST-Double-Net model, trained and tested on this dataset, are shown in Figure 6 for four magnification levels, where the values along the main diagonals represent the counts of correctly classified images. Based on these confusion matrices and formulae (16)–(19), the values of the evaluation metrics, achieved by the proposed model for each subclass at each magnification level, are shown in Tables 5–8.

At magnification level of 40×, the ST-Double-Net values of all metrics reached 100% in five subclasses, proving that the model performs exceptionally well in the classification of A, MC, PC, PT, and TA subclasses. At the same time, its recall value in the F subclass also reached 100%. However, in the LC subclass, its precision and F1-score only reached 80.56% and 86.57%, respectively (much lower than the other subclasses), which could be attributed to the high similarity of samples between subclasses.

At magnification level of 100×, the proposed model achieved excellent classification results in the A and TA subclasses, with all metrics reaching 100%. At the same time, its precision in the MC and PC subclasses also reached 100%. However, its classification performance in the LC subclass was low again – reaching only 87.88% for precision, 85.29% for recall, and 86.57% for F1-score.

At magnification level of 200×, the proposed model scored 100% for all metrics in two subclasses, proving that the model performs exceptionally well in the classification of A and PC. At the same time, its recall value in the MC and TA subclasses also reached 100%.

At magnification level of 400×, all metric values of ST-Double-Net reached 100% for the A and TA subclasses, and precision for the MC subclass also reached 100%.

Overall, the proposed ST-Double-Net model achieved very good results across all subclasses at all magnification levels.

TABLE 5. Multi-class classification performance results of ST-Double-Net at 40× magnification on BreaKHis dataset.

Subclass	Evaluation Metrics (%)			
	Accuracy	Precision	Recall	F1-score
A	**100**	**100**	**100**	**100**
DC	97.47	98.80	95.35	97.04
F	99.75	98.04	**100**	99.01
LC	97.72	80.56	93.55	86.57
MC	**100**	**100**	**100**	**100**
PC	**100**	**100**	**100**	**100**
PT	**100**	**100**	**100**	**100**
TA	**100**	**100**	**100**	**100**

The overall accuracy and the average precision, recall, and F1-score of the proposed ST-Double-Net model, achieved on the BreaKHis dataset, are shown in Table 9. It can be seen that compared with only using the Swin Transformer to extract global feature information for multi-class classification (c.f., Table 3), the fusion of global and local feature information in ST-Double-Net allows to significantly improve its multi-class classification performance. More specifically, compared with only using the Swin Transformer, the overall accuracy of ST-Double-Net has increased by 1.52∼4.25 percentage points, the average precision has increased by 1.19∼6.77 percentage points (with a small drop of 0.57 percentage points at magnification level of 400×), the average recall has increased by 2.86∼5.28 percentage points, and the average F1-score has increased by 2.05∼6.02 percentage points.

Next, on the same dataset (BreaKHis), we compared the overall accuracy of the proposed ST-Double-Net model to that of state-of-the-art models, based on their results reported in the corresponding literature sources, as summarized in Table 10. From this table, it is evident that the proposed model

TABLE 6. Multi-class classification performance results of ST-Double-Net at 100× magnification on BreaKHis dataset.

Subclass	Evaluation Metrics (%)			
	Accuracy	Precision	Recall	F1-score
A	**100**	**100**	**100**	**100**
DC	97.58	96.20	98.33	97.25
F	99.52	98.08	98.08	98.08
LC	97.83	87.88	85.29	86.57
MC	99.76	**100**	97.73	98.85
PC	99.52	**100**	92.86	96.30
PT	99.52	95.83	95.83	95.83
TA	**100**	**100**	**100**	**100**

TABLE 7. Multi-class classification performance results of ST-Double-Net at 200× magnification on BreaKHis dataset.

Subclass	Evaluation Metrics (%)			
	Accuracy	Precision	Recall	F1-score
A	**100**	**100**	**100**	**100**
DC	98.00	97.24	98.32	97.78
F	99.50	98.08	98.08	98.08
LC	98.00	92.86	81.25	86.67
MC	99.75	97.50	**100**	98.73
PC	**100**	**100**	**100**	**100**
PT	99.50	95.24	95.24	95.24
TA	99.75	96.55	**100**	98.24

TABLE 8. Multi-class classification performance results of ST-Double-Net at 400× magnification on BreaKHis dataset.

Subclass	Evaluation Metrics (%)			
	Accuracy	Precision	Recall	F1-score
A	**100**	**100**	**100**	**100**
DC	95.88	95.78	95.21	95.50
F	99.45	97.73	97.73	97.73
LC	96.43	65.22	75.00	69.77
MC	99.45	**100**	93.75	96.77
PC	99.45	96.43	96.30	96.30
PT	99.45	96.30	96.30	96.30
TA	**100**	**100**	**100**	**100**

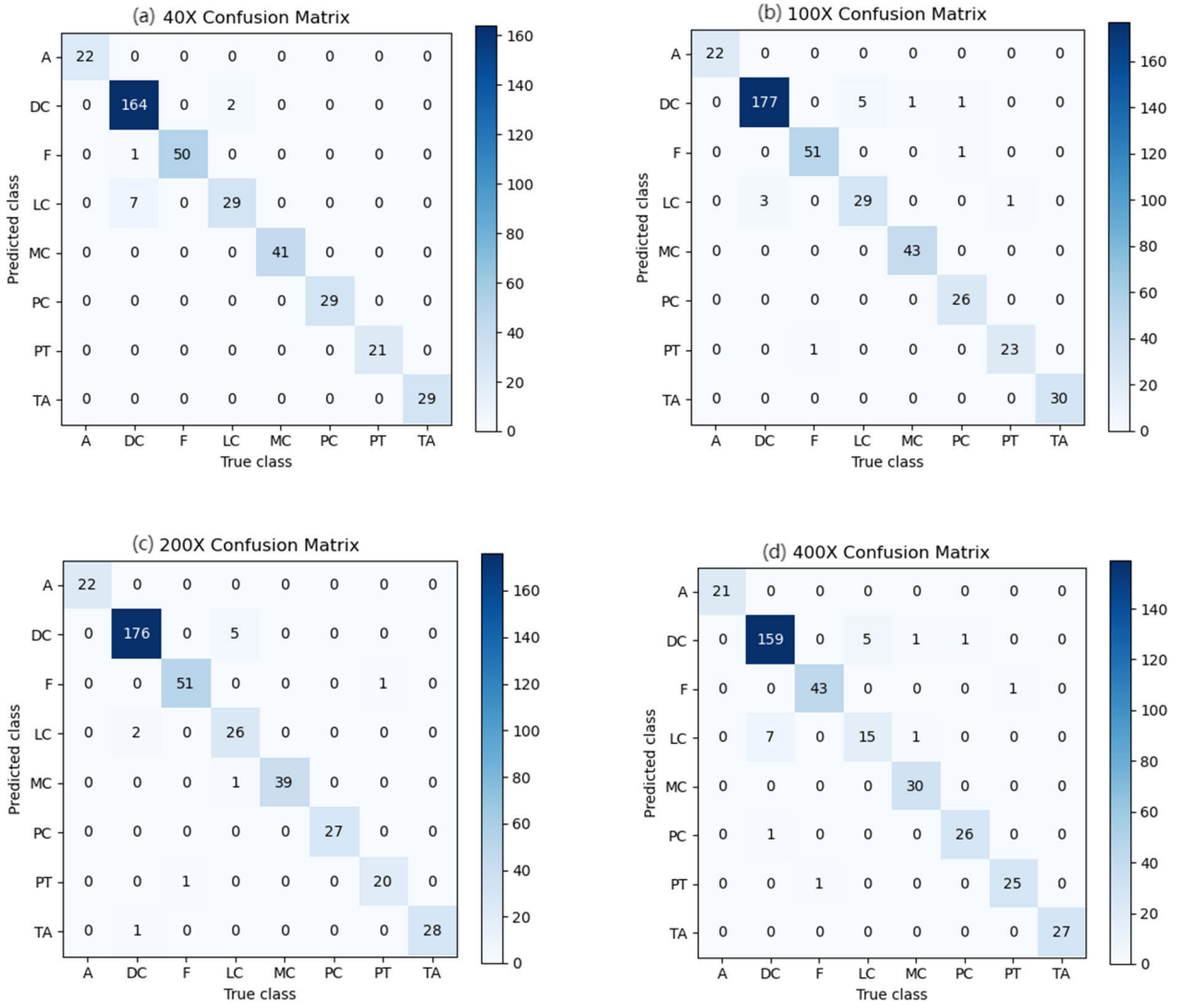


FIGURE 6. The multi-class confusion matrices of ST-Double-Net on BreaKHis dataset at various magnification levels: (a) 40×, (b) 100×, (c) 200×, and (d) 400×.

TABLE 9. Overall/Average multi-class classification performance of the proposed model at various magnification levels on BreaKHis dataset.

Model	Magnification	Evaluation Metrics (%)			
		Overall accuracy	Average precision	Average recall	Average F1-score
ST-Double-Net (*proposed*)	40 ×	**97.47**	97.18	**98.61**	**97.83**
	100 ×	96.86	**97.25**	96.02	96.61
	200 ×	97.25	97.18	96.61	96.84
	400 ×	95.05	93.93	94.29	94.05

TABLE 10. Overall accuracy of the proposed model vs. state-of-the-art models at various magnification levels on BreaKHis dataset.

Model	Overall accuracy (%) at different magnification levels			
	40 ×	100 ×	200 ×	400 ×
PGLCM-SARF [39]	93.88	93.97	94.57	94.77
BreaST-Net [40]	96.00	92.60	93.50	93.40
CNN-LSTM [41]	96.30	92.60	88.04	92.51
PGLCM-IBL [42]	89.49	85.52	85.85	88.54
CNN-SVM [43]	91.00	89.00	88.00	86.00
E7WVAD [44]	93.80	93.40	93.30	91.80
ST-Double-Net (*proposed*)	**97.47**	**96.86**	**97.25**	**95.05**

outperforms all state-of-the-art models at all magnification levels.

2) ON BACH DATASET

The confusion matrix of the proposed model ST-Double-Net, trained and tested on this dataset, is shown in Figure 7, where the values along the main diagonal represent the numbers of correctly classified images. Based on this confusion matrix and formulae (16)–(19), the values of the evaluation metrics, achieved by the proposed model for each class, are shown in Table 11. As can be seen, ST-Double-Net performed well on

all four classes, even achieving 100% precision for the in-situ carcinoma class, and 100% recall for the normal class.

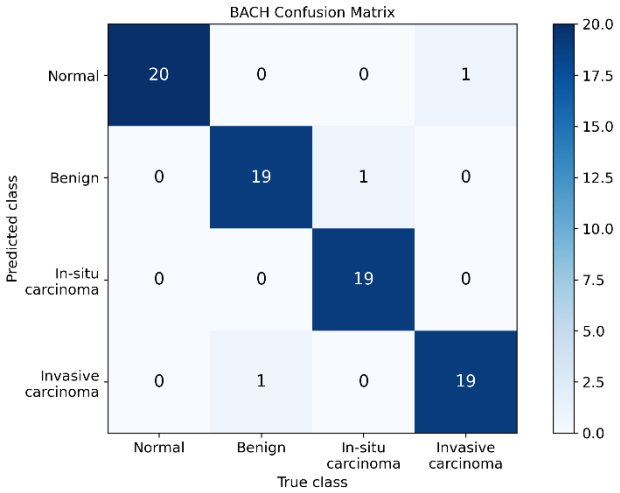


FIGURE 7. The multi-class confusion matrix of ST-Double-Net on BACH dataset.

TABLE 11. Multi-class classification performance results of ST-Double-Net on BACH dataset.

Class	Evaluation Metrics（%）			
	Accuracy	Precision	Recall	F1-score
Normal	**98.75**	95.24	**100**	**97.56**
Benign	97.50	95.00	95.00	95.00
In-situ carcinoma	**98.75**	**100**	95.00	97.44
Invasive carcinoma	97.50	95.00	95.00	95.00

Next, on the same dataset (BACH), we compared the overall accuracy of the proposed ST-Double-Net model to that of state-of-the-art models, based on their results, reported in the corresponding literature sources, as summarized in Table 12. From this table, it is evident that the proposed model is superior to all state-of-the-art models.

TABLE 12. Overall accuracy of the proposed model vs. state-of-the-art models on BACH dataset.

Model	Overall accuracy (%)
Unsupervised model [45]	87.20
HACT-NET [46]	91.00
Sujatha et al. [47]	92.00
MDF-FNet [48]	86.00
Sarker et al. [49]	96.00
FCCS-Net [50]	91.25
ST-Double-Net (*proposed*)	**96.25**

F. ABLATION STUDY

We also conducted ablation study experiments with different components of the proposed model, results of which are shown in Tables 13 and 14, respectively for the BreaKHis and

TABLE 13. Ablation study results on BreaKHis dataset.

Model's components	Magnification	Evaluation metrics（%）			
		Overall accuracy	Average precision	Average recall	Average F1-score
Swin Transformer (base)	40 ×	95.95	95.99	95.75	95.78
	100 ×	93.72	93.65	91.69	92.46
	200 ×	93.00	90.41	91.33	90.82
	400 ×	92.52	**94.50**	89.75	91.38
Swin Transformer + two stages	40 ×	96.46	95.86	97.69	96.60
	100 ×	93.96	92.74	95.36	93.74
	200 ×	94.75	93.97	95.18	94.46
	400 ×	94.49	92.34	**95.33**	93.53
Swin Transformer + two stages + HMC (i.e., ST-Double-Net *proposed*)	40 ×	**97.47**	**97.18**	**98.61**	**97.83**
	100 ×	**96.86**	**97.25**	**96.02**	**96.61**
	200 ×	**97.25**	**97.18**	**96.61**	**96.84**
	400 ×	**95.05**	93.93	94.29	**94.05**

TABLE 14. Ablation study results on BACH dataset.

Model's components	Evaluation metrics（%）			
	Overall accuracy	Average precision	Average recall	Average F1-score
Swin Transformer	92.50	92.55	92.50	92.50
Swin Transformer + two stages	95.00	95.12	95.00	95.03
Swin Transformer + two stages + HMC (i.e., ST-Double-Net *proposed*)	**96.25**	**96.31**	**96.25**	**96.25**

BACH datasets. In the second step of this study, after building a two-stage model based on the Swin Transformer, the values of all evaluation metrics improved on both datasets, except for precision at magnification levels of 40×, 100× and 400× on the BreaKHis dataset. In the third step, after adding the HMC module to complete the proposed model, all metrics reached top values on both datasets, except for precision and recall at magnification level of 400× on the BreaKHis dataset.

V. CONCLUSION

This paper has proposed a two-stage fine-grained classification model based on weakly supervised target localization. Specifically, in response to the problem of easily confused target objects and backgrounds in breast cancer pathological images, the use of a newly designed heatmap cropping (HMC) module with weakly supervised target positioning has been proposed. By using the feature extraction capability of Swin Transformer, the proposed model does not require other additional information such as labeling boxes. The classification network in the first stage is focused on the lesion area in the images; the area is cropped, thereby achieving extraction of fine-grained *global* features. In the second-stage classification network, Swin Transformer is used to extract *local* features that are difficult to extract by other traditional networks. Finally, the local feature map and the global feature map are fused in the channel dimension and input into

the classification head to complete the final classification. Results, obtained on two public datasets, have demonstrated the superiority of the proposed two-stage classification model to state-of-the-art models.

However, the proposed ST-Double-Net model has also some limitations. The model relies on weakly supervised target localization techniques to extract lesion areas, which may be limited by the accuracy of target localization. If the target localization is not accurate, it may lead to inaccurate feature extraction, which affects the final classification performance.

In view of the small differences between the classes of pathological images themselves, in the future, we plan to develop and integrate an attention module as to improve the performance of the proposed model.

REFERENCES

[1] L. Sun and X. Yong, "Research on classification method of mammography based on deep learning," *Comput. Eng. Appl.*, vol. 54, no. 21, pp. 13–19, 2018.

[2] R. M. Mann, N. Cho, and L. Moy, "Breast MRI: State of the art," *Radiology*, vol. 292, no. 3, pp. 520–536, Sep. 2019.

[3] Y.-L. Huang, Y.-R. Jiang, D.-R. Chen, and W. K. Moon, "Level set contouring for breast tumor in sonography," *J. Digit. Imag.*, vol. 20, no. 3, pp. 238–247, Sep. 2007.

[4] J. V. Horvat, D. M. Keating, H. Rodrigues-Duarte, E. A. Morris, and V. L. Mango, "Calcifications at digital breast tomosynthesis: Imaging features and biopsy techniques," *RadioGraphics*, vol. 39, no. 2, pp. 307–318, Mar. 2019.

[5] 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.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[7] N. C. Kurian, A. Sethi, A. R. Konduru, A. Mahajan, and S. U. Rane, "A 2021 update on cancer image analytics with deep learning," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 4, Jul. 2021, Art. no. e1410.

[8] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.

[9] M. Radeta, R. Freitas, C. Rodrigues, A. Zuniga, N. T. Nguyen, H. Flores, and P. Nurmi, "Man and the machine: Effects of AI-assisted human labeling on interactive annotation of real-time video streams," *ACM Trans. Interact. Intell. Syst.*, vol. 14, no. 12, pp. 1–22, 2024.

[10] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[12] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3512–3520.

[13] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016.

[14] G. Aresta et al., "BACH: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122–139, Aug. 2019.

[15] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, and I. Soerjomataram, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *Breast*, vol. 66, pp. 15–23, Dec. 2022.

[16] Y. S. Vang, Z. Chen, and X. Xie, "Deep learning framework for multi-class breast cancer histology image classification," in *Proc. 15th Int. Conf. Image Anal. Recognit.*, Povoa de Varzim, Portugal. Cham, Switzerland: Springer, Jun. 2018, pp. 914–922.

[17] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," *Expert Syst. Appl.*, vol. 120, pp. 103–115, Apr. 2019.

[18] D. A. Ragab, O. Attallah, M. Sharkas, J. Ren, and S. Marshall, "A framework for breast cancer classification using multi-DCNNs," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104245.

[19] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100154.

[20] Z. Jiang, Z. Dong, J. Fan, Y. Yu, Y. Xian, and Z. Wang, "Breast TransFG plus: Transformer-based fine-grained classification model for breast cancer grading in hematoxylin-eosin stained pathological images," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105284, doi: 10.1016/j.bspc.2023.105284.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[22] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: A systematic survey," Rensselaer Polytech. Inst., Troy, NY, USA, Tech. Rep. TR-05-09, 2005.

[23] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *J. Med. Imag.*, vol. 5, no. 3, Jul. 2018, Art. no. 036501.

[24] K. Song, X.-S. Wei, X. Shu, R.-J. Song, and J. Lu, "Bi-modal progressive mask attention for fine-grained recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 7006–7018, 2020.

[25] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," 2014, *arXiv:1406.2952*.

[26] X. Ke, Y. Huang, and W. Guo, "Weakly supervised fine-grained image classification via two-level attention activation model," *Comput. Vis. Image Understand.*, vol. 218, Apr. 2022, Art. no. 103408.

[27] Y. Zhang, K. Jia, and Z. Wang, "Part-aware fine-grained object categorization using weakly supervised part detection network," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1345–1357, May 2020.

[28] A. Patil, D. Tamboli, S. Meena, D. Anand, and A. Sethi, "Breast cancer histopathology image classification and localization using multiple instance learning," in *Proc. IEEE Int. WIE Conf. Electr. Comput. Eng. (WIECON-ECE)*, Nov. 2019, pp. 1–4.

[29] M. Fan, T. Chakraborti, E. I.-C. Chang, Y. Xu, and J. Rittscher, "Microscopic fine-grained instance classification through deep attention," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Lima, Peru. 2020. Cham, Switzerland: Springer, Oct. 2020, pp. 490–499.

[30] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler, "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8296–8307.

[31] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 8778–8788.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

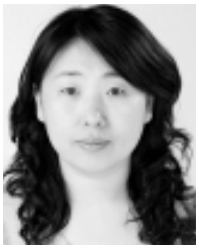
[33] A. Ashurov, S. A. Chelloug, A. Tselykh, M. S. A. Muthanna, A. Muthanna, and M. S. A. M. Al-Gaashani, "Improved breast cancer classification through combining transfer learning and attention mechanism," *Life*, vol. 13, no. 9, p. 1945, Sep. 2023, doi: 10.3390/life13091945.

[34] Y. Zhou, C. Zhang, and S. Gao, "Breast cancer classification from histopathological images using resolution adaptive network," *IEEE Access*, vol. 10, pp. 35977–35991, 2022, doi: 10.1109/ACCESS.2022.3163822.

[35] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. 19th Int. Conf. Comput. Statist.*, Paris, France. Heidelberg, Germany: Springer, Aug. 2010, pp. 177–186.

[36] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[37] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[39] Y. Wang, X. Deng, H. Shao, and Y. Jiang, "Multi-scale feature fusion for histopathological image categorisation in breast cancer," *Comput. Methods Biomech. Biomed. Eng., Imag. Vis.*, vol. 11, no. 6, pp. 2350–2362, Nov. 2023.
[40] S. Tummala, J. Kim, and S. Kadry, "BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of Swin transformers," *Mathematics*, vol. 10, no. 21, p. 4109, Nov. 2022.
[41] M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, "Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning," *BMC Med. Imag.*, vol. 23, no. 1, pp. 1–15, Jan. 2023.
[42] J. Li, J. Shi, H. Su, and L. Gao, "Breast cancer histopathological image recognition based on pyramid gray level co-occurrence matrix and incremental broad learning," *Electronics*, vol. 11, no. 15, p. 2322, Jul. 2022.
[43] C. S. Vikranth, B. Jagadeesh, K. Rakesh, D. Mohammad, S. Krishna, and A. S. R. Ajai, "Computer assisted diagnosis of breast cancer using histopathology images and convolutional neural networks," in *Proc. 2nd Int. Conf. Artif. Intell. Signal Process. (AISP)*, Feb. 2022, pp. 1–6.
[44] H. Zerouaoui, O. E. Alaoui, and A. Idri, "New design strategies of deep heterogenous convolutional neural networks ensembles for breast cancer diagnosis," *Multimedia Tools Appl.*, vol. 83, no. 24, pp. 65189–65220, Jan. 2024.
[45] T. S. Sheikh, J.-Y. Kim, J. Shim, and M. Cho, "Unsupervised learning based on multiple descriptors for WSIs diagnosis," *Diagnostics*, vol. 12, no. 6, p. 1480, Jun. 2022.
[46] P. Pati, G. Jaume, A. Foncubierta-Rodríguez, F. Feroce, A. M. Anniciello, G. Scognamiglio, N. Brancati, M. Fiche, E. Dubruc, D. Riccio, M. Di Bonito, G. De Pietro, G. Botti, J.-P. Thiran, M. Frucci, O. Goksel, and M. Gabrani, "Hierarchical graph representations in digital pathology," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102264.
[47] R. Sujatha, J. M. Chatterjee, A. Angelopoulou, E. Kapetanios, P. N. Srinivasu, and D. J. Hemanth, "A transfer learning-based system for grading breast invasive ductal carcinoma," *IET Image Process.*, vol. 17, no. 7, pp. 1979–1990, May 2023.
[48] C. Xu, K. Yi, N. Jiang, X. Li, M. Zhong, and Y. Zhang, "MDFF-Net: A multi-dimensional feature fusion network for breast histopathology image classification," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107385.
[49] M. M. K. Sarker, F. Akram, M. Alsharid, V. K. Singh, R. Yasrab, and E. Elyan, "Efficient breast cancer classification network with dual squeeze and excitation in histopathological images," *Diagnostics*, vol. 13, no. 1, p. 103, Dec. 2022.
[50] R. Maurya, N. N. Pandey, M. K. Dutta, and M. Karnati, "FCCS-Net: Breast cancer classification using multi-level fully convolutional-channel and spatial attention-based transfer learning approach," *Biomed. Signal Process. Control*, vol. 94, Aug. 2024, Art. no. 106258.



SHENGNAN HAO received the B.S. degree from North China University of Science and Technology, China, in 1996, and the M.S. degree from Beijing University of Technology, China, in 2009. In 1996, she joined North China University of Science and Technology, and became an Associate Professor, in 2009. Her current research interests include complex systems, impulsive systems, and stochastic control.

YIHAN JIA was born in 2001. She received the bachelor's degree in engineering from Hebei Science and Technology Normal University, in 2022. She is currently pursuing the master's degree with North China University of Science and Technology. Her research interests include computer vision and medical image processing.

JIANUO LIU was born in 2000. She received the B.S. degree from North China University of Science and Technology, in 2022, where she is currently pursuing the master's degree. Her research interests include machine vision and graphic image processing.

ZHIWU WANG received the Ph.D. degree from Tianjin Medical University, in 2014. He is currently a Chief Physician with the Department of Radiotherapy and Chemotherapy, Tangshan People's Hospital. He is engaged in the comprehensive medical treatment of lung cancer and digestive system tumors. His current research interests include medical data processing and screening tumor immunotherapy effect prediction markers based on artificial intelligence.

CHUNLING LIU received the Ph.D. degree from the Chinese Academy of Medical Sciences, Peking Union Medical College, in 2014, specializing in cell research. She is currently an attending Physician with the Pathology Department, Tangshan People's Hospital.

ZHANLIN JI (Member, IEEE) received the Ph.D. degree from the University of Limerick, Ireland, in 2010. Currently, he is a Professor with Zhejiang Agriculture and Forestry University, China; and an Associate Researcher with the Telecommunications Research Centre (TRC), University of Limerick, Ireland. He was recipient of the Irish Research Council for Science, Engineering and Technology (IRCSET) Post-Graduate Research Scholarship, in 2008; and an IRC Postdoctoral Fellowship, in 2013. He has authored/co-authored more than 100 research papers in refereed journals and conferences. His research interests include the ubiquitous consumer wireless world (UCWW), Internet of Things (IoT), cloud computing, big data management, and artificial intelligence (AI)-based image processing.

IVAN GANCHEV (Senior Member, IEEE) received the Engineering and Ph.D. degrees (summa cum laude) from Saint-Petersburg University of Telecommunications, in 1989 and 1995, respectively. He is an International Telecommunications Union (ITU-T) Invited Expert and an Institution of Engineering and Technology (IET) Invited Lecturer. He is currently affiliated with the University of Limerick, Ireland; the University of Plovdiv "Paisii Hilendarski," Bulgaria; and the Institute of Mathematics and Informatics—Bulgarian Academy of Sciences (IMI-BAS), Bulgaria. He was involved in more than 40 international and national research projects. He has authored/co-authored one monographic book, three textbooks, four edited books, and more than 300 research papers in refereed international journals, books, and conference proceedings. He has served on the TPC of more than 400 prestigious international conferences/symposia/workshops. He is on the editorial board of and has served as a guest editor for multiple prestigious international journals.