MSHT: Multi-Stage Hybrid Transformer for the ROSE Image Analysis of Pancreatic Cancer

Tianyi Zhang, Yunlu Feng, Yu Zhao, Guangda Fan, Aiming Yang, Shangqing Lyu, Peng Zhang ©, Fan Song, Chenbin Ma ©, Yangyang Sun, Youdan Feng, and Guanglei Zhang ©, Member, IEEE

Abstract—Pancreatic cancer is one of the most malignant cancers with high mortality. The rapid on-site evaluation (ROSE) technique can significantly accelerate the diagnostic workflow of pancreatic cancer by immediately analyzing the fast-stained cytopathological images with on-site pathologists. However, the broader expansion of ROSE diagnosis has been hindered by the shortage of experienced pathologists. Deep learning has great potential for the automatic classification of ROSE images in diagnosis. But it is challenging to model the complicated local and global image features. The traditional convolutional neural network (CNN) structure can effectively extract spatial features, while it tends to ignore global features when the prominent local features are misleading. In contrast, the Transformer structure has excellent advantages in capturing global features and long-range relations, while it has limited ability in utilizing local features. We propose a multi-stage hybrid Transformer (MSHT) to combine the strengths of both, where a CNN backbone robustly extracts multi-stage local features at different scales as the attention guidance, and a Transformer encodes them for sophisticated global modeling. Going beyond the strength of each single method, the MSHT can simultaneously enhance the Transformer global modeling ability with the local guidance from CNN features. To evaluate the method in this unexplored field, a dataset of 4240 ROSE images is collected where MSHT achieves 95.68% in classification accuracy with more accurate attention regions. The distinctively superior results compared to the state-of-the-art models make MSHT extremely promising for cytopathological image analysis.

Index Terms—Cytopathology, deep learning, pancreatic cancer, rapid on-site evaluation (ROSE), Transformer.

I. INTRODUCTION

Pancreatic cancer is one of the most highly malignant tumors of the digestive system, with a very low 5-year survival rate of about 10% [1]. Due to the lack of symptoms and the shortage of proper screening technology, patients diagnosed with pancreatic cancer generally present in the advanced stage, which brings striking challenges for the treatment of the patients [2]. Therefore, if diagnosed earlier before the metastasis of cancer cells, it would have a better prognosis and a higher survival rate [3].

Endoscopic ultrasonography-guided fine-needle aspiration (EUS-FNA) has achieved a remarkable diagnostic accuracy for pancreatic cancer and is now widely used [4]. With the utilization of endoscopic devices and sampling techniques, the current diagnostic sensitivity of pancreatic cancer is 90% to 98% and the specificity is 95% to 100% [5], [6]. The relatively low diagnostic sensitivity is mainly caused by the inadequacy of the pancreatic pathology sampling. To reduce the risk of complications and the pain of the patient during puncture surgery, the times of the needle punctures are generally limited, which may lead to the missed diagnosis. Such limitations encourage researchers to make efforts toward the improvement of the sampling procedure [7].

Rapid on-site evaluation (ROSE), which refers to the real-time cytopathological evaluation during the FNA procedure, has been widely used with the expectation to decrease the diagnostic period with fewer needle punctures and increase the sample adequacy [8]. The availability of ROSE may reduce the inadequate sample rate of pancreatic cancer by 10%-18% [9] and thus avoid the missed diagnosis during the EUS-FNA procedure. However, the core difficulty has been pointed out to be limited pathologist staffing in [10] and [11] which reported that the ROSE was available in only 48% and 55% of European and Asian centers. In comparison, it is available in nearly 98% of the USA centers. To expand ROSE diagnosis worldwide, automating the workflow is highly in need.

With the development of computer-aided diagnosis (CAD) techniques and the increase of computing power, artificial intelligence has played an important role in health care. The analysis of cytopathological images with deep learning technology has been widely reported with promising diagnostic accuracy in analyzing breast cancer, cervical cancer, and gastric cancer [12], [13].
Although deep learning technology shows excellent potential in the diagnosis of pancreatic cancer along with the clinical innovation ROSE technique, very few works have been reported in recent years.

In the ROSE images, the normal cells distinguish from their cancerous counterparts in various local features and their global distribution features. Firstly, the size and shape features are considered in identifying cancerous cell changes, but these features are not decisive alone, without comparing the background red cells, etc. Secondly, the distribution of cells is prominent in general cases, where normal cells are distributed regularly, while the cancer cells are more discrete, and their distances to each other are imparity. Thirdly, because of the tridimensional structure of the cell smear, the normal cell clusters follow a trend of the same direction in their orientation, or it is a smooth transition. In contrast, the cancerous clusters show irregular distribution. However, deciding the correct categories of samples is challenging due to the complex local and global features changing with background perturbations. Besides, the features are relatively inconspicuous in some cases compared with other pathological tasks.

A possible deep-learning-based based CAD system can be based on image classification. The first CAD system for ROSE images of EUS-FNA specimen was introduced by Hashimoto et al. [14] in 2018, in which the authors created a stage-by-stage deep learning system that achieved the sensitivity, specificity, and accuracy all at 80% on a private dataset of 450 ROSE images. In 2020, the following work by the same group [15] achieved the accuracies of 93% and 89% in diagnosing pancreatic cancer with two deep learning models called ImageNet-CNNbn and RetinaNet on a larger dataset of 1440 ROSE images. In general, the limited accuracy of the related works may be linked to the following reasons. On the one hand, the scarcity of the ROSE images due to the time-consuming workflow and the scarcity of experienced pathologists both make dataset construction an arduous task. On the other hand, ROSE images usually contain complex backgrounds, such as the presence of noisy areas and perturbations like red cells, fibers, and vacuoles, which demand a robust feature extractor for better analysis [13].

Both of the related works by Hashimoto [14], [15] applied convolutional neural networks (CNN) in the analysis of ROSE images. The early stage of CNN usually has more local attention biases, capturing local features and patterns. However, due to the limited kernel size of CNN, only the spatial related pixels can be considered in a single layer. This character is beneficial to the generalization ability of the classification task, but it may limit the global modeling ability in analyzing ROSE images. Meanwhile, although the reception field of CNN models could be made larger when designed with deeper layers, the networks still easily focus on local features due to their inductive bias instead of global features [12]. Thus, achieving an optimized classification performance for ROSE images is challenging.

Attention-based methods shed light on the improvement of deep learning since the introduction of Transformer [16]. To enhance the global modeling ability, the encoder-based Transformer was introduced into the field of computer vision by [17]. Due to its global receptive field and long-distance modeling strength, it outperforms a series of state-of-the-art CNN models [18], [19], [20], [21], [22], [23]. The utilization of Transformer blocks in the analysis of medical images was gradually studied. Frontier work GasHis-Transformer [24], which was based on Inception [20] and BoT [25] networks, has achieved a state-of-the-art result of 96.8% in accuracy for diagnosing gastric cancer. In the research of cytopathological image analysis with Transformer structure, cell-DETR [26], which was based on DETR [27] structure, was introduced for instance segmentation on cells. However, the Transformer-based models require a large-scale dataset to constrain and fully perform their self-attention ability (Dosovitskiy et al. [17]), which makes the combination of Transformer networks with ROSE image analysis of pancreatic cancer a challenging task.

To address the issue mentioned above and construct a classification model with remarkable accuracy for ROSE image analysis, the hybrid idea between the Transformer blocks and the CNN blocks was introduced in this study. When the CNN backbone goes deeper, different stages provide various grained features of the pancreatic cells, which encode the abstractive information through different scales. Additionally, the CNN structure provides a robust feature-extracting method by its small parameter design and the character of inductive bias, which could achieve better convergence on the limited dataset. Furthermore, due to the spatial difference among the cells and their global distribution features, which are the morphological characteristics to distinguish cancer cells from normal ones, the self-attention mechanism of the Transformer is introduced to process the extracted spatial features at a global scale. The difficulty falls on how to encode the features from CNN in the global modeling process of the Transformers, in which the focus guided decoder (FGD) structure is specially designed for converting multi-stage CNN features into multi-stage attention guidance. With such hybrid construction, CNN can provide a robust fine-grained feature extractor towards the perturbation of noisy areas existing in the ROSE images and provide supportive features across different scales. At the same time, the attention guidance enables the Transformer structure to globally model the relevant information of long-distance regions without missing prominent local features.

In general, we proposed a multi-stage hybrid Transformer (MSHT) model with the integration of the CNN backbone and focus guided decoder (FGD) structure. The CNN is firstly used to generate different scales of features from the ROSE images, representing the abstractive spatial features of the cells. Then the focus blocks of the FGD structure encode the CNN feature maps into feature sequences to carry the attention information from early stages and use them as the attention guidance. Moreover, the Transformer decoder of the FGD structure is designed to globally model the deeper patterns of the feature sequences by its global receptive field, which sophisticatedly decodes the spatial attention biases into the global modeling process. Finally, a multi-layer perceptron (MLP) is used to classify the images by taking out the class token in the Transformer outputs.

The contributions of the MSHT are as follows:

- As one of the first works in ROSE image analysis, the proposed MSHT firstly introduces the cutting-edge Transformer blocks from the computer vision field, which improves the classification performance by the enhanced global modeling ability of the Transformer.
- Taking full advantage of the unique multi-stage hybrid architecture of the Transformer and CNN, MSHT is empowered to be able to extract multi-level features.
Fig. 1. The architecture of the proposed Multi-stage Hybrid Transformer (MSHT) model for the classification of ROSE images of pancreatic cancer. (a) The architecture of MSHT. (b) The focus block of the FGD structure. (c) The decoder of the FGD structure. MHSA denotes multi-head self-attention, MHGA denotes multi-head guided-attention, LN denotes layer norm block, FFN denotes the feed-forward network, and MLP denotes multi-layer perceptron.

across different scales and thus can achieve much higher classification accuracy by using various grained features and comprehensive distribution patterns of the pancreatic cells.

- To leverage the spatial attention biases from CNN in the Transformer global modeling process, the FGD focus block is designed to integrate spatial features as the guidance in global modeling. The MSHT is evaluated on the most extensive ROSE dataset to our knowledge, which presents the remarkable performance and solid interpretability of our method. Meanwhile, this innovation could shed light on cytopathology and histopathology, where cells and tissues share distinctive features by their spatial changes and global relevance.

II. METHODS

The proposed Multi-stage Hybrid Transformer (MSHT) is designed for the analysis of cytopathological ROSE images of pancreatic cancer, which aims to diagnose pancreatic cancer faster and without time occupation of the pathologists. Toward the characters of ROSE images, the hybrid structure design of MSHT combines the robust feature-extracting strength of CNN on spatial features, and the global modeling power of Transformer on the abstractive patterns. As shown in Fig. 1, in MSHT, the CNN backbone robustly generates different grained feature maps from different stages. Then, the focus block is specially designed based on locational consistency to encode the CNN features, which extracts the locational attention biases of different stages of the CNN backbone. Lastly, through the explicitly designed FGD structure, the stage-wise attention information is fused into the Transformer global modeling process.

A. Backbone CNN: ResNet50

The MSHT model uses the ResNet50 [19] as our backbone to extract features robustly from the input ROSE images. The ResNet50 is designed by stacking a Stem block and four stages of CNN bottleneck blocks, where the blocks downsample the image $I \in \mathbb{R}^{3 \times E \times E}$ with edge size of $E$ into abstractive features. To effectively fuse the multi-stage features from the CNN backbone with the Transformer decoders, we modified the original backbone ResNet50 by taking out the feature maps of $l \in \{1, 2, 3, 4\}$ stages and connecting them to the focus blocks of the FGD structure. Structurally, the calculation flow is reserved to maintain the mainstream feature extraction, while the last stage feature maps serve as the first inputs of the FGD structure for more complex global modeling.

B. Embedding Modules and Focus Block

1) Embedding Module and Hybrid Embedding Block: As shown in Fig. 1(a), at the beginning of the FGD structure, the hybrid embedding block transposes the feature maps of the last CNN stage $l = 4$ into feature patches for global modeling using the Transformer blocks. The hybrid embedding block is designed to bridge the last-stage features between the CNN and Transformer in different stages. Fig. 2 illustrates the calculation
The architecture of the hybrid embedding block, which is composed of a convolutional patch division process and an embedding process.

During the patch projection process, a single CNN layer is used to split the input feature maps into \( N = E^2 / p^2 \) patches by setting its kernel size and stride, equal to the same patch size \( p \). To maintain the most features in the hybrid process, the minimal final stage embedding patch size \( p = 1 \) are adopted. Meanwhile, the channel setting of the CNN layer needs to project the dimension of the CNN feature maps \( C_t \) to match the input dimension of the Transformer \( D \). After the CNN projection, the feature maps are transformed into feature sequences through the embedding module to meet the prerequisites of the Transformer. The class token and positional encoding are used in all embedding processes, which is hired in both the hybrid embedding block and the focus block.

The class token \( x_{class} \) is a learnable matrix that serves as the first token in Vision Transformers (Dosovitskiy et al. [17]). In MSHT, a global shared parameter patch \( x_{class} \in \mathbb{R}^{1 \times D} \) is deployed to carry the classification information throughout the FGD structure, as shown in Fig. 1(a). Initially, it is an empty token represented by a zero tensor \( x_{class} \in \mathbb{R}^{1 \times D} \). In the final stage of FGD, only this token is connected to the classification MLP so that the classification information can be encoded to it during the forward process.

Positional encoding is a process that allows the model to observe the location information. In the positional encoding, a standard learnable design is followed in which \( N+1 \) of \( D \) dimensional parameter is randomly initiated as \( M_{pos} \in \mathbb{R}^{(N+1) \times D} \). Through the backpropagation process, the positional information can be encoded into the data processing workflow.

In the embedding process of the embedding module, feature patches are flattened and transposed into tokens \( [x_{1,1}^t; x_{1,2}^t; \ldots; x_{N,1}^t] \). Then, the class token \( x_{class} \) is concatenated as the first input token, and the one-dimensional learnable positional encoding is added. With these designs, the feature maps of the last stage \( l = 4 \) are transformed into \( N + 1 \) tokens, and each token has the exact dimension \( D \). The output of the hybrid embedding process is \( z_0 \).

\[
z_0 = [x_{class}; x_{1,1}^t; x_{1,2}^t; \ldots; x_{N,1}^t] + M_{pos} \tag{1}
\]

2) Focus Block: In each backbone stage \( l \in \{1, 2, 3, 4\} \), the focus block transfers the early feature map \( x_l \in \mathbb{R}^{C_l \times E_l \times E_l} \) (with the corresponding edge size of \( E_l \) and the channel size of \( C_l \)) into the deeper stages as attention guidance which delivers the early-stage features such as textures. As shown in Fig. 1(b), the focus block of the FGD structure is composed of 3 main steps:

(a) Attention module for processing the CNN features
(b) Dual pooling path for gathering the attention biases inside the feature maps of each stage
(c) Embedding module for transforming the attention biases into feature sequence of the Transformer decoder.

Firstly, the attention module used in the FGD focus block obtains the output feature maps \( x_l^p \in \mathbb{R}^{C_l \times E_l \times E_l} \) by processing the attention features on \( x_l \in \mathbb{R}^{C_l \times E_l \times E_l} \) from different CNN backbone stages \( l \). To obtain the attention activation from local stages from backbone, the SimAM block [28] is specially applied. SimAM aims to identify discriminative neurons in deep neural networks based on findings from neuroscience, thus enhancing feature maps with a neuron-based attention mechanism. In this parameter-free attention module, the spatial bias of each early-stage neuron \( t \) is captured by its minimal activation energy \( e_t^a \).

\[
e_t^a = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{2}
\]

where \( \lambda \) is a hyperparameter, \( \hat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \hat{\mu})^2 \) and \( \hat{\mu} = \frac{1}{M} \sum_{i=1}^{M} x_i \) are calculated by the feature map \( x_l \in \mathbb{R}^{C_l \times E_l \times E_l} \), \( M = E_l \times E_l \) is the number of the energy functions in each channel \( C_l \), and each \( x_i \) is one of the surrounding neurons of the target neuron \( t \). The lower the \( e_t^a \) is, the more prominent the neuron \( t \) acts. Therefore, the attention feature maps \( x_l^p \in \mathbb{R}^{C_l \times E_l \times E_l} \) are calculated in (3), where \( \hat{E}_l \) groups all \( e_t^a \) across channel and spatial dimensions following [28].

\[
x_l^t = \text{sigmoid} \left( \frac{1}{\hat{E}_l} \right) \odot x_l \tag{3}
\]

Secondly, a dual attention-gathering strategy is proposed to filter the spatial attention activation for the global modeling process through the stages. Specifically, to obtain general and prominent features, feature maps \( x_l^p \in \mathbb{R}^{C_l \times E_l \times E_l} \) after the attention module are considered by the design of two parallel pooling paths, including the max pooling and average pooling in (4) and (5) as follows:

\[
x_l^{\text{Maxpool}} = \text{MaxPooling(Attention}(x_l)) \tag{4}
\]

\[
x_l^{\text{Avgpool}} = \text{AvgPooling(Attention}(x_l)) \tag{5}
\]

Inspired by the gaze and glance process in human eyes, both pooling strategies share the same pooling window size \( p_l \) and transform the feature maps into the feature patches with the patch size \( p_l \). In each stage’s focus block, we adopt different pooling window sizes \( p_l = E_l \times E_l \) to align the feature maps to the same size of the last stage (when \( l = 4 \)). The outputs containing spatial attention information can be represented by \( x_l^{\text{Maxpool}} \in \mathbb{R}^{C_l \times (E_l/p_l) \times (E_l/p_l)} \) and \( x_l^{\text{Avgpool}} \in \mathbb{R}^{C_l \times (E_l/p_l) \times (E_l/p_l)} \).

Thirdly, the two separate convolution layers share the same input and output channel sizes of \( C_l \) and \( D \), and with their kernel size and stride of 1 are applied to alter the dimension of the feature maps, as shown in (6) and (7).

\[
f_l^{\text{Maxpool}} = \text{CNN1}(x_l^{\text{Maxpool}}) \tag{6}
\]

\[
f_l^{\text{Avgpool}} = \text{CNN2}(x_l^{\text{Avgpool}}) \tag{7}
\]

After the 2 CNN layers, feature patches \( f_l^{\text{Maxpool}} \in \mathbb{R}^{D \times (E_l/p_l) \times (E_l/p_l)} \) and \( f_l^{\text{Avgpool}} \in \mathbb{R}^{D \times (E_l/p_l) \times (E_l/p_l)} \) are transformed into \( f_l^{p_l} \in \mathbb{R}^{(E_l/p_l)^2 \times D} \) and \( f_l^{p_l} \in \mathbb{R}^{(E_l/p_l)^2 \times D} \) by flattening
and transposing action, as shown in (8) and (9).

\[
f^q_l = \text{Transpose}(\text{Flatten}(f^{\text{Maxpool}}_{l})) \quad (8)
\]

\[
f^k_l = \text{Transpose}(\text{Flatten}(f^{\text{Avgpool}}_{l})) \quad (9)
\]

Lastly, the same embedding strategy in Fig. 2 is used inside the FGD focus block of each stage \( l \) to obtain embedded guidance sequences \( q_l \in \mathbb{R}^{(N+1) \times D} \) and \( k_l \in \mathbb{R}^{(N+1) \times D} \). These sequences will act as query and key in Transformer modeling.

\[
q_l = \text{Concatenate}(x_{\text{class}}, f^q_l) + M_{\text{pos}} \quad (10)
\]

\[
k_l = \text{Concatenate}(x_{\text{class}}, f^k_l) + M_{\text{pos}} \quad (11)
\]

C. Focus Guided Decoder (FGD)

1) FGD Decoder: In the mainstream of the hybrid structure FGD, the global modeling process takes advantage of local attention features. To align the encoded attention information from different stages of CNN to Transformer feature sequences, the 4-decoder stacking structure is designed as the primary branch in the FGD structure, as shown in Fig. 1(a) and (c). The input of the first decoder is the feature patches \( z_0 \in \mathbb{R}^{(N+1) \times D} \) processed by the hybrid embedding block, which is connected to the last stage of the CNN backbone in (12). The outputs of each decoder have the same size of \( (N + 1) \times D \), and the information flow is then transmitted to the next decoder stage. Each FGD decoder is stacked with the multi-head self-attention (MHSA), Multi-head Guided Attention (MHGA), and Feed-Forward Network (FFN) blocks. The decoders can perform more robustly under different conditions with the pre-norm strategy and the residual connection.

In the workflow of each decoder, as shown in Fig. 1(c), a MHSA block (13) is firstly used to process the information from the last CNN block. As a typical Transformer block, its self-attention structure achieves the purpose of long-distance feature modeling and gathers abstractive information globally. After the MHSA block, the FFN block (14) stabilizes the processing workflow, by introducing learnable projection layers. Based on the design of a 2-layer MLP with the connection of non-linear activation units, FFN can also support the desire for generalization ability.

The focus block takes advantage of the inductive biases from different stages of the CNN backbone and encodes them as attention guidance. These features can be fused with the features of global modeling by the MHGA attention mechanism in (15). Finally, after the MHGA block, we use residual connection and FFN block (16) to stabilize the training process in the same way as the original Transformer decoder design [16]. The pre-norm strategy is applied to stabilize the gradient during the training process, therefore LayerNorm (LN) module is inserted before the MHSA, MHGA, and FFN blocks.

Toward the modeling of the abstractive patterns, the decoders are connected to each other by the stacking design. At the last stage of the FGD structure in (17), only the class token \( x^\text{output}_{\text{class}} \) of the output sequences is connected to the MLP classification head.

\[
z_0 = \text{hybrid\_Embedding}(\text{CNN\_backbone}(I)) \quad (12)
\]

\[
z^1_l = \text{MHSA}(\text{LN}(z^1_{l-1})) + z^{1}_{l-1}, l = 1, 2, 3, 4 \quad (13)
\]

\[
z^2_l = \text{MLP}(\text{LN}(z^1_l)) + z^1_l, l = 1, 2, 3, 4 \quad (14)
\]

2) Multi-head Self-attention (MHSA): The Transformer models are known for the long-distance modeling ability of MHSA [16], which contributes to its global modeling process. As shown in Fig. 3, MHSA has the tremendous advantages of capturing embedded feature maps which achieve its strength in the global modeling process of the FGD structure.

In the MHSA block, the given feature patches \( f_l \) of the stage \( l \) are transformed into three sequences called \( q^\text{self}_l, k^\text{self}_l, v^\text{self}_l \) by the scale inner dot product with three learnable matrices \( w_q, w_k, w_v \) as \( q^\text{self}_l = w_q \cdot f_l / \sqrt{D}, k^\text{self}_l = w_k \cdot f_l / \sqrt{D}, v^\text{self}_l = w_v \cdot f_l / \sqrt{D} \). The \( q^\text{self}_l, k^\text{self}_l, v^\text{self}_l \) indicates the query, key, and value for the attention process. After the transform, the resize calculation is applied to partition them into \([q^1_l, q^2_l, \ldots, q^N_l], [k^1_l, k^2_l, \ldots, k^N_l]\) and \([v^1_l, v^2_l, \ldots, v^N_l]\) where \( h \) is the number of heads. The self-attention operation is proceeded in each head \( h \), for a given embedded patch \( f_l, f^h_l = \text{SoftMax}(k^h_l f^h_l + b^h_l)^T \cdot v^h_l \). The output \([f^1_l, f^2_l, \ldots, f^N_l]\) is resized back to the feature dimension \( D \) to remove the additional head dimension and the \( f^\text{MHSA}_l \in \mathbb{R}^{N \times D} \) can be obtained.

3) Multi-head Guided-attention (MHGA): As shown in Fig. 4, different from the MHSA, the MHGA block aims to process attention sequences from the FGD focus blocks. The attention information captured by the dual information flow in the FGD focus block plays a significant role in MHGA, which are used to encode additional spatial information biases within the global modeling process of the FGD decoder blocks.

As the first step, the inputs query and key \( (q_l, k_l) \) which contain prominent and general attention information from early CNN stages are transformed into feature patches \( q^\text{guide}_l = w_q \cdot f^q_l / \sqrt{D} \) and \( k^\text{guide}_l = w_k \cdot f^k_l / \sqrt{D} \). Then, the scaled inner dot
product and SoftMax layers are designed as the MHSA structure. After the resizing operation, which splits the feature patches into different heads, \( v^h \) is multiplied with \( \text{SoftMax}(q^h)^T \cdot k^h \) to obtain \( G^h \) in each head. Lastly, the multi-head information is combined by rearranging and resizing operations for the output \( f^h \).

### D. Multi-stage Hybrid Transformer

The proposed MSHT model is a hybrid model that integrates CNN with Transformer. The main feature extraction workflow and the intermediate outputs of the backbone CNN are considered within the multi-stage design. As we are analyzing a fine-grained task of cell image classification, we use the input image size \( E = 384 \) to maintain the feature scales in the modeling. The feature maps are downsampled through the backbone CNN stages \( l \in [1, 2, 3, 4] \) with the size of \( 256 \times 96 \times 96 \) (stage 1 \( C_1 = 256, E_1 = 96 \)), \( 512 \times 48 \times 48 \) (stage 2 \( C_2 = 512, E_2 = 48 \)), \( 1024 \times 24 \times 24 \) (stage 3 \( C_3 = 1024, E_3 = 24 \)), and \( 2048 \times 12 \times 12 \) (stage 4 \( C_4 = 2048, E_4 = 12 \)).

After the feature extracting process with the CNN backbone, the hybrid embedding block is introduced to transform the feature maps from the last CNN stage into the input feature patches of the Transformer blocks. The features are embedded from \( 2048 \times 12 \times 12 \) to \( N \times D \), which is \( 144 \times 768 \) (as the last stage \( l = 4 \) embedding size \( p \) is 1 and \( N = E_l^2 / p^2 \), and the embedding dimension \( D \) is set to 768 following the Transformer design convention). At the end of the embedding block, the size of the output feature patches is expanded from \( 144 \times 768 \) to \( 145 \times 768 \), due to the concatenation of an empty class token \( x_{\text{class}} \) with the size of \( 1 \times 768 \).

Connecting closely to the different stages of the CNN backbone, each focus block firstly downsamples the feature maps from different stages to the same edge size as the last feature map (size of \( E_4 = 12 \)) and then embeds the attention-guided information into two feature sequences \( q_l \in \mathbb{R}^{(N+1) \times D} \) and \( k_l \in \mathbb{R}^{(N+1) \times D} \) with the same size of \( 145 \times 768 \). As shown in (13–16), the global modeling process by the Transformer decoders can model the embedded sequences with attention guidance sequences. After the four stages of the FGD structure, the first dimension \( x_{\text{output}}^{\text{class}} \) which carries the classification information in the output sequences, is connected to the MLP classification head.

Lastly, the MLP combined with the SoftMax layer projects the feature dimension \( D \) of the class token \( x_{\text{output}}^{\text{class}} \) (size of 768) to the category indexes representing the predicted confidence on each class.

### III. EXPERIMENT AND RESULTS

#### A. Dataset

The ROSE diagnosis was implemented in Peking Union Medical College Hospital (PUMCH), and the data were collected under the supervision of senior pathologists. The 4240 pancreatic cytopathological samples were separately obtained from cyto-pathology analysis, in which 4240 images in total were independently sampled after the EUS-FNA surgery and diff-quik giemsa staining procedure of ROSE examinations. The enrolled images were performed by two microscope digital cameras (Basler ScA1 and Olympus DP73) with Olympus BX53 and Nikon Eclipse Ci-S microscopes. From 2019 to 2021, a total amount of 1518 pancreatic cancer images and 2722 normal pancreatic cell images were collected, and the classification labels were confirmed by senior pathologists of PUMCH. The images were saved in ‘.jpg’ format with resolutions of 1390×1038, 2400×1800, and 5480×3648 under the same magnification of 400 times during the data sampling process. Since all input images should share the same length-to-width ratio, they were resized to 1390×1038 to maintain the same magnification factor in the range.

#### B. Experimental Setting

To comprehensively and objectively evaluate the models on the ROSE images, we adopted a 5-fold training setting. Firstly, we randomly divided the 4240 individual images into two groups: a training-validation set and an independent test set with a ratio of 8:2. Then the training-validation set was randomly divided into five datasets, representing 5-folds with approximately the same number of images in each fold. In the fold \( k \) (\( k \in \{1, 2, \ldots, 5\} \)), the k fold dataset was used as the validation dataset, while the data from the remaining four folds were used as the training datasets. In each experiment, the model was trained five times individually, with different validation and training datasets each time, but the independent test dataset was shared. In each epoch, the model was trained and tested on the training and validation datasets. The model for a given epoch, which showed the best accuracy on the validation dataset, was saved as the output model of the fold \( k \).

A certain online data-augmentation strategy was implied in the experiments to recreate views under the microscope, as shown in Fig. 5. During each training process, the input images were randomly rotated, and a center area with a size of 700×700 pixels was cropped as the input data. Then, the random horizontal flip and random ‘color-jitter’ (with the setting of brightness = 0.15, contrast = 0.3, saturation = 0.3, and hue = 0.06) were used to recreate the white balance shifting and other perturbations during sampling. In training, all augmentations were sequentially triggered. In the validation and testing process, only the ‘CenterCrop’ operation was used to reserve the central pixels inside the 700×700 pixels boundary.

In both training and validation processes, the images were lastly resized to 384×384 and transformed into tensors. The Adam optimizer was used in training with a learning rate of 1e-5.
and a momentum of 0.05. The cosine learning rate decay strategy was adopted to reduce the learning rate four times sequentially in the training process. In all folds of the experiments, a batch size setting of 8 images was used, and 50 epochs were trained and tested on the training and validation datasets. The counterparts of the MHST model were trained with the same hyperparameter setting or better-performed hyperparameters. All experiments used a same random seed for reproducibility.

The experiments were carried out and recorded online on the Google Colab pro+ platform. In each experiment, a 16 GB Nvidia P100-PCIe GPU was used with Python version 3.7.12 and Pytorch version 1.9.0+cu111. The model was built based on the Pytorch [29] and timm library [30]. To obtain better results on the ROSE datasets, transfer learning was applied in all models with their official model weights. The proposed MSHT model was pre-trained on the ImageNet-1k [39] for 150 epochs with a batch size of 210 on 3 Nvidia A6000 GPUs. The pre-training learning rate was set to 1e-6 and with a cosine learning rate decay for ten times. The implementation details along with the entire experiment scripts and records have been released online. Our implemented with Colab notebook are repeatable with the same settings.

C. Evaluation Criteria

The experimental results of MHST and its counterparts were measured by Accuracy (Acc), Precision (Pre), Recall (Rec), Sensitivity (Sen), Specificity (Spe), Positive predictive value (PPV), Negative predictive value (NPV), and F1-score. During the measurement of the 2-class classification task (positive or negative), the criteria were calculated by the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), all of the detailed results were recorded to indicate their research values. To reveal the interpretability of the models, the gradient-weighted class-activation-mapping (Grad-CAM) [31] method was used to visualize the attention regions.

D. MSHT for ROSE Analysis

In Table I, several evaluation criteria are used to assess the experimental result, and the 5-fold average result for each indicator is presented to indicate the overall performance of MSHT and its counterparts. In the 5-fold training process, the MSHT achieves an average of 97.53%, 95.38%, 96.54%, and 98.08% for Acc, Spe, Sen, PPV, and NPV. Meanwhile, the average indicators during the validation process are 94.37%, 96.69%, 90.20%, 93.93%, and 94.67%. In the independent test dataset, 95.68% of images are correctly classified, and MSHT achieved 96.95%, 93.40%, 94.54%, and 96.35% for Spe, Sen, PPV, and NPV. For average F1-score, MSHT achieves 96.55%, 91.98% and 93.94% in training, validation and testing process. The results of MSHT are significantly better than the early works [14], [15] (with 93% and 89% accuracy), and MSHT is fully validated on a larger dataset of 4240 images compared with their 1440 images.

The interpretability is essential, and we evaluate the attention regions of the model based on the identification of cells and cell clusters. Since the distribution is complex, it is very challenging to identify the cells when they are not nearby. However, the related information is important in ROSE. Therefore, the three key points are summarized in the ROSE analysis of attention regions: 1) The identification of pancreatic and cancer cells. 2) The identification of the background (attention should not be on the background cells and empty places). 3) The identification of the trend of cells and cell clusters (global relationships in different situations).

The results indicate that the MSHT has achieved encouraging results on the ROSE images, as shown in Fig. 6, presenting solid interpretability through the visualization of its attention regions by the Grad-CAM technique. In Fig. 6, four typical cases are illustrated with predictive confidence, and heatmaps represent

### Table I: Results of the Counterpart Models on the ROSE Dataset

| Model         | F1-score (%) | ACC (%) | SPE (%) | SEN (%) | PPV (%) | NPV (%) | MSHT significance p-value (t-test) |
|---------------|--------------|---------|---------|---------|---------|---------|-----------------------------------|
| EfficientNet b3 [22] | 90.51 (0.86) | 93.29 (0.68) | 95.48 (1.66) | 89.37 (2.14) | 91.80 (2.70) | 94.19 (1.00) | 0.0007                |
| MobileNet V3 [23]     | 90.80 (0.98) | 93.44 (0.76) | 95.11 (1.21) | 90.43 (0.86) | 91.20 (1.97) | 94.70 (0.44) | 0.0024                |
| Inception V3 [20]     | 91.49 (0.75) | 93.84 (0.54) | 94.49 (0.48) | 92.67 (1.09) | 90.35 (0.78) | 95.86 (0.59) | 0.0036                |
| Xception [21]         | 92.55 (0.63) | 94.69 (0.48) | 96.07 (0.83) | 92.21 (0.85) | 92.91 (1.34) | 95.68 (0.44) | 0.0216                |
| VGG-19 [18]           | 92.78 (0.68) | 94.83 (0.59) | 96.03 (1.91) | 92.67 (2.25) | 93.02 (2.91) | 95.96 (1.14) | 0.0586                |
| VGG-16 [18]           | 92.95 (0.34) | 94.92 (0.26) | 95.66 (1.42) | 93.60 (2.32) | 92.42 (2.19) | 96.44 (1.20) | 0.0720                |
| ResNet50 [19]         | 93.12 (0.75) | 95.02 (0.39) | 95.51 (1.18) | 94.13 (1.15) | 92.17 (1.85) | 96.70 (0.60) | 0.0761                |
| ViT [17]              | 92.37 (0.99) | 94.50 (0.36) | 95.26 (0.47) | 93.14 (0.69) | 91.63 (0.77) | 96.16 (0.26) | 0.0131                |
| DeiT [34]             | 92.42 (0.90) | 94.52 (0.55) | 95.04 (0.71) | 93.60 (2.60) | 91.34 (0.97) | 96.41 (1.35) | 0.0275                |
| Swin Transformer [35] | 93.03 (1.10) | 94.92 (0.90) | 95.18 (1.88) | 94.46 (1.54) | 91.74 (2.88) | 96.87 (0.81) | 0.1116                |
| Conformer [36]        | 92.16 (1.36) | 94.52 (0.83) | 96.80 (1.04) | 90.43 (3.75) | 94.11 (1.63) | 94.84 (1.90) | 0.0598                |
| Cosavit [37]          | 89.77 (1.09) | 92.33 (0.76) | 94.52 (0.64) | 88.38 (1.57) | 89.99 (1.10) | 93.60 (0.80) | 0.0004                |
| Mobvlievit [38]       | 90.48 (1.28) | 93.48 (0.72) | 97.21 (0.63) | 86.80 (3.01) | 94.58 (1.05) | 93.00 (1.41) | 0.0102                |
| **MSHT (Ours)**       | **93.94 (0.97)** | **95.68 (0.76)** | **96.95 (1.52)** | **93.40 (0.83)** | **94.54 (2.51)** | **95.36 (0.41)** | --                     |
the decision boundaries of the feature regions. Within each heatmap generated by the Grad-CAM analysis, the red areas indicate where the specific classification output is highly correlated, while the blue areas are correlated with other categories. In case 1, two typical negative samples are shown with clear background and complex surroundings. The attention regions indicate the MSHT has a better preference on each cell area, which proves the hybrid structure has clearly captured the long-distance distribution and spatial features. In case 2, MSHT has evidently identified the pancreatic cells without the redundant attention spent, even in the noisy surroundings. In cases 3 and 4, MSHT clearly identifies cancerous cells even though there were fewer positive samples than negative samples due to the imbalance of the dataset. The spatial difference of cancer cells and their distinctive distribution are correctly captured by the MSHT, as the attention regions indicate. The accurate attention regions prove the robust interpretation of the proposed MSHT and imply its clinical potential.

Additionally, the problem of misclassification that needs to be overcome is given by taking two examples, as shown in Fig. 7. Three of the five 5-fold MSHT models have misclassified a specific negative sample to the positive condition. By the analysis of senior pathologists, the reason lies in the fluctuation of the squeezed sample. The nuclei of the pancreatic cells in the false positive sample present a sawtooth shape, and the distance to each other are no-longer smooth transition, which misleads MSHT by these cancerous patterns. A few positive samples are misclassified as negative ones. In this typical false positive sample, the cell number is very limited, so the information on their arrangement and relative size are not supportive as the other samples. Moreover, the trend of cell clusters and the size distributions are unclear, which can mislead junior pathologist. Compared with senior pathologists, this limited information makes it difficult for MSHT to distinguish cancer cells.

### E. Comparisons With Counterpart Models

Due to the limited work of this field, we evaluated the proposed MSHT with dedicatedly optimized counterpart models. As a promising option in the cytopathology research, the seven state-of-the-art CNNs included: VGG-16, VGG-19 [18], EfficientNet_b3 [22], ResNet50 [19], Inception-V3 [20], Xception [21] and MobileNet-V3 [23].

As the first work to introduce the Transformer into ROSE image analysis, three cutting-edge Transformer-based models (vision transformers) from the computer vision field were used as the comparison models, including ViT [17], DeiT [34], and Swin Transformer [35]. Three recent hybrid computer vision backbones combining CNN and Transformers were also compared, including Conformer [36], Crossvit [37], and Mobilevit [38]. The transfer learning technique was used for all models with the official weight of models pre-trained on the ImageNet [39]. The models were compared with the same criteria on the test dataset after converging at the same or better-optimized hyperparameter settings.

Table 1 shows the 5-fold average results of models on the testing process of ROSE image classification, which indicates that the proposed MSHT achieves the highest overall results with Acc (95.68%) and F1-score (93.94%), which outperforms its counterparts significantly. In the experiments, most of the counterparts achieve approximate Acc less than 95%, while only ResNet reaches 95.02%, slightly higher than 95%. Although the imbalance of the dataset (1518 positive images vs. 2722 negative images) increases the difficulty for models to recognize positive samples, compared with other models, MSHT has achieved significantly higher results in Spe and PPV which contribute to the higher overall performance in Acc and F1-scores.

The results indicate that the pure CNN or transformer-based models are less effective in the ROSE classification task than MSHT. In Table 1, two CNNs and two Transformers achieve higher Sen and NPV than MSHT. The optimized VGG-16 and DeiT models achieve slightly higher Sen (+0.20%, +0.20%) and NPV (+0.09%, +0.06%) compared with MSHT. The best-performed CNN counterpart ResNet50 achieves 0.73% higher Sen and 0.34% higher NPV than MSHT, but it shows obvious lower Spe (-1.43%) and PPV (-2.37%) results. Furthermore, the best-performed Transformer model, Swin Transformer achieves higher in Sen (+1.06%) and NPV (+0.52%) with much lower results in Spe (-1.77%) and PPV (-2.80%). Since the negative samples are more than the positive ones, higher Sen and NPV are easier to obtain than other criteria. Such results indicate that the high Acc performed counterparts share the same biases on the negative samples.

Compared with pure CNN or Transformer models, most hybrid models achieve higher Spe and PPV. Although Mobilevit achieves slightly higher (+0.26% and 0.04%) results in these two indicators, MSHT achieves much higher overall F1-score performance (+3.46%). In the AI-aided diagnosis process for ROSE images, Spe and PPV are more important than the opposite indicators Sen and NPV. This is because the misdiagnosis of the AI system leads to the inadequate sampling of pancreatic tissues and further results in reducing the final diagnostic accuracy. In contrast, the missed diagnosis usually leads to more sampling of pancreatic tissues and does not affect the final diagnostic accuracy. The results indicate that the hybrid methods are more suitable for ROSE than pure structure design, especially when the MSHT performs better overall.

The MSHT outperforms the rest counterparts by higher results in all six criteria F1-score, Acc, Spe, Sen, PPV, and NPV across the 5-folds. We use a one-tail t-Test (paired two samples for means) to measure the significance of the increase in F1-score. Statistically, MSHT achieves significantly better results than most of its counterparts except Swin Transformer, where the p-value is not less than 0.1. The results may be linked to the high variance of the performance of the Swin Transformer, which is less robust on ROSE, and its general results are much lower (-0.90% in F1-score).

In terms of interpretability, as shown in Fig. 8, the models perform differently when their attention regions are visualized by the Grad-CAM technique. In general, the CNNs tend to focus on contiguous regions covering both cells and unrelated backgrounds, while the Transformers tend to focus on discrete regions representing the global features. The CNN models show acceptable attention regions when dealing with negative samples, while they focus on both cells and background when facing positive samples. Under noisy conditions, as shown in...
negative sample 1, the CNNs tend to suffer from background perturbation, which is not ideal in clinical applications. In contrast, Vision Transformers show different biases. Briefly, the Transformers can capture the features of the pancreatic cells as they have clear attention regions on them. In negative sample 2, the attention regions are discrete while the cells are close to each other, which indicates the limitation of spatial related modeling in the Transformers. The Transformers easily focus on the background of the positive samples, indicating that the models have not learned the correct patterns due to the limited samples.

The MSHT outperforms CNNs as its attention regions cover the distinctive cells in most cases. Furthermore, the MSHT shows its robust focus in noisy situations. Compared with the Transformers, in addition to being able to focus on cells globally, its attention regions are more aggregated following the trend of cells. The MSHT performs delightfully in the interpretability study, which can correctly distinguish the samples and focus on the cell groups like the senior pathologists.

More cases with different background complexity are illustrated in Fig. 9, which are identified correctly by MSHT, even when the surroundings of cells may confuse junior doctors. Moreover, the classification confidence (obtained by a SoftMax calculation of the model predication) is decisive in most cases, which indicates its performance is reliable.

As the MSHT is a multi-stage hybrid strategy towards CNN and Transformer, we illustrate the attention map of each CNN stage and the corresponding Transformer stage with Grad-CAM. The results illustrated in Fig. 10 prove the effectiveness of the local-to-global attention guidance design. In the early CNN stages, the attention regions tend to focus on spatial areas. In the early Transformer stages, they cannot focus on the whole area of the global distribution. From the CNN to the Transformer stages, the attention maps with spatial focus gradually increase the sharpness of the attention regions in the global modeling. At the last stage of MSHT, the attention regions are ideal, with a balance of prominent spatial features and global identification.

F. Ablation Studies of MSHT

To evaluate the structural improvement and explore the efficiency of the proposed MSHT, a series of ablation studies were drawn out under the same training and validation setting. Except for the pre-training experiment, MSHT ablation counterparts had their backbone weights initiated by the official ResNet50, and the weights of the remaining structures were initialized randomly for a fair comparison.

1) The Effectiveness of MSHT FGD Structure: To evaluate the effectiveness of the FGD structure, we designed two ablation models with different hybrid structures. Firstly, the Hybrid1 model was designed to reveal the significance of the proposed multi-stage hybrid structure. In Hybrid1, by a stacking design of four stages of ResNet50 [19] structure and eight Transformer encoder modules, the same calculation mainstream as the proposed MSHT was recreated. Meanwhile, the 3-stage multi-stage
Fig. 10. Examples of stage-wise attention visualization. Through the early CNN stages to deep Transformer stages, the attention maps with spatial focus gradually increase the sharpness of the attention regions in the global modeling.

hybrid design could improve the resolution of the CNN feature maps, although more calculation was required, due to its feature map expansion (edge size increased from 12 to 24). Therefore, we designed the Hybrid3 to compare the proposed FGD structure in terms of the stage depth and the size of the feature maps.

As shown in Table II, the proposed 4-stage design of the MSHT (Hybrid2_No_PreTrain) model achieves the best performance in terms of Acc, Sen, NPV, and F1-score with 95.30%, 93.66%, 96.47%, and 93.45%. Meanwhile, the Hybrid1 model achieves the same result in Sen, but it only reaches 94.90% Acc and 92.93 F1-score in general criteria. The results indicate a distinctive gap in the absence of the FGD structure. The stacking design may limit the pattern modeling ability in the deeper layers, as the mainstreams of the model are the same in the Hybrid1 and MSHT (Hybrid2_No_PreTrain) design.

Taking account of the complexity of FGD, the standard 4-stages of ResNet structure has smaller feature sizes and higher dimensions carrying deeper CNN features, which is used in MSHT. The 3-stage Hybrid3 model reaches the highest Spe and PPV with 96.54% (+0.33%) and 93.66% (+0.38%), which are slightly better than the 4-stages design model. Since the Hybrid3 model only reaches 94.73% Acc (-0.57%) with more calculation cost in the experiments, it can be seen that the number of stages of FGD plays a crucial role. These results indicate that the FGD structure is effective and that the 4-stage design achieves the best results cost-efficiently.

2) The Effectiveness of Class Token and Positional Encoding: Following the computer vision studies, the class token and positional encoding are pivotal to reveal the performance of the Transformer based model. In MSHT, the global sharing class token and positional encoding are designed to work as the messenger throughout the FGD structure. Therefore, the Hybrid2_No_CLS_Token and Hybrid2_No_Pos_emb are designed specially, without using the class token and positional encoding in the FGD structure.

As shown in Table II, the results indicate that the MSHT design achieves evidently higher performance in all six criteria. Without the class token, the model Hybrid2_No_CLS_Token reaches 94.85% Acc and 92.77% F1-score, which are slightly higher than the performance of Hybrid2_No_Pos_emb with 94.71% and 92.56% only. The results prove that class token design can influence the outcome, but the positional encoding is more critical to the model, as the knowledge of the distribution of the cells contributes to the global modeling process of the Transformer.

3) The Effectiveness of Attention Module in Focus Block: The proposed hybrid design requires representative local attention information. Therefore, an attention module is adopted in the FGD focus block before the dual pooling layers to extract the attention activation from local stages. The focus block equipped with the widely used CBAM [32] module and SE [33] module are compared as Hybrid2_CBAM and Hybrid2_SE. Noticeably, the attention module SimAM deployed in the focus block is a parameter-free structure that reduces the model size and complexity, the Hybrid2_No_ATT is therefore designed to prove its effectiveness.

The proposed MSHT with SimAM module achieves a higher result in all six criteria compared with no attention module design and CBAM attention module version. Compared with no attention design Hybrid2_No_ATT, un-pre-trained MSHT (Hybrid2_No_PreTrain) achieves 0.78% higher Acc and 1.07%
higher F1-score. Only a slightly higher difference can be shown compared with Hybrid2_CBAM in general criteria, Hybrid2_No_PreTrain achieves 0.19% and 0.25% higher in Acc and F1-score. The Hybrid2_SE achieves slightly higher (by +0.04%) in a specific indicator Spe than Hybrid2_No_PreTrain. The proposed MSHT achieves the ambition of feature converting in FGD with a cost-efficient module SimAM, which achieves the best general result at the lowest calculation cost.

IV. DISCUSSION AND CONCLUSION

Through the stage-wise hybrid design, the proposed MSHT introduces the cutting-edge Transformer structure into ROSE image analysis and performs more sophisticated results than the backbone CNNs. Together with the CNN feature extraction structure, the introduction of the Transformer module improves the global modeling capabilities by globally integrating the local features of cells and the relationships between cells. In the early stages of MSHT, the local features extracted by convolution kernels can be related to specific positions due to the CNN structure. While in the deeper stages, the global modeled feature sequences obtained by the Transformer self-attention mechanism, which contains more abstractive information, the misalignment may occur if the features are combined directly. As a hybrid design, the MSHT takes locational consistency as the key to attention encoding.

To encode the features and their positions into the global pattern modeling process, two steps are designed in the FGD structure. Firstly, to obtain local attention guidance sequences \( q_l \) and \( k_l \), the representative spatial features are enhanced by the parallel converting design in the focus block. Therefore, the attention-based structure can capture the local CNN features. Secondly, the locally obtained sequences \( q_l \) and \( k_l \) are converted to global guidance weight by the scale inner dot product in MHGA. Their locational attention is encoded to transformer global modeling sequence \( f_l^{\text{MHGA}} \in \mathbb{R}^{N \times D} \). This design avoids the misalignment between the Transformer global abstractive sequences and the encoded CNN feature sequences by encoding the attention bias as a locational enhancement. The attention visualization proves the strength of this design, as the MSHT obtains better attention regions on both prominent local areas and the global distributional features.

Compared with the state-of-the-art models, MSHT achieves significantly higher Acc and F1-score. At the feature extracting stages, the MSHT model inherits the strength of the inductive bias of the CNNs blocks, which contributes to robust performance under complex surroundings. Benefiting from the global modeling design of attention mechanism, MSHT sophisticatedly captures global distributive features by introducing the Transformer module. The hybrid structure plays an important role, which subtly extracts the attention information from each stage of the CNN backbone to guide the global modeling process. Compared with the pure attention-mechanism-based models ViT and DeiT, MSHT performs higher overall results with the multi-stage hybrid structure converting attention guidance from the CNN backbone.

Using the early-stage feature maps of CNN, the attention biases are obtained by the FGD structure to guide the process of global modeling. Two different global sharing designs, class token and positional encoding carrying the through-out information are used in each embedding process, leading to a more significant improvement than other ablation models. The FGD structure influences the effectiveness of the Transformer blocks and CNN blocks. Compared with the directly stacking strategy without FGD structure, feature maps of different CNN stages are used as the attention biases, which directly contribute to the modeling process of Transformer blocks. Focusing on the number of stages in the CNN backbone, deeper stage (four stages vs. three stages) design enhances the stability and robustness when the model is faced with small medical datasets.

Data scarcity is pivotal in performing its full potential in medical image analysis. From a data-related view of data-augmentation, which is meant to recreate the clinical sampling condition, the generalizability of MSHT can be increased. More conclusively, however, to alleviate the limitation in the training process of MSHT, the transfer learning strategy is hired following the intention of general image understanding. To obtain better results on the ROSE datasets, all models are initialized by transfer learning strategy in the experiments. In Table II, comparing with un-pre-trained MSHT model, the proposed MSHT achieves distinctive higher Acc (+0.38%), Spe (+0.74%), PPV (+1.26%), and F1-score (+0.49%). The overall criteria Acc and F1-score are significantly higher, but the data-driven Sen and NPV are slightly lower than the randomly initialized counterparts.

For the complicated characters of ROSE images, MSHT can extract the local features at the cell level and then models the global long-distance relationship of cells. As a trending clinical technique, ROSE can save patients by introducing the on-site workload. MSHT classifies cancerous and normal pancreatic cells with high accuracy and sheds light on a possible AI-based CAD system. The application of such systems can significantly boost the diagnosis process and aid the great demand for the workforce of on-site pathologists. The inspiring results enable AI-on-site diagnosis, which can further benefit broader populations where on-site pathologists are rare.

In conclusion, the proposed MSHT model achieves the state-of-the-art classification performance on the cytopathological ROSE image analysis via its unique multi-stage hybrid architecture, which can effectively combine the advantages of CNN and Transformer. The MSHT model, together with the clinically innovative ROSE technology, facilitates the exploration of rapid and pathologist-free EUS-FNA procedures and expands its potential for broader application in pancreatic cancer diagnosis.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Colorectal cancer statistics, 2020,” CA: A Cancer J. Clinicians, vol. 70, no. 3, pp. 145–164, 2020.
[2] J. D. Mizrahi, R. Sarana, J. W. Valle, and R. T. Shroff, “Pancreatic cancer,” Lancet, vol. 395, 2020, Art. no. 10242.
[3] S. P. Pereira et al., “Early detection of pancreatic cancer,” Lancet Gastroenterol. Hepatol., vol. 5, no. 7, pp. 698–710, 2020.
[4] M. Kitan, T. Yoshida, M. Itouga, T. Tamura, K. Harakawa, and Y. Yamashita, “Impact of endoscopic ultrasonography on diagnosis of pancreatic cancer,” J. Gastroenterol., vol. 54, no. 1, pp. 19–32, 2019.
[5] H. Matsuyoshi et al., “Endoscopic ultrasonography-guided fine needle aspiration for the diagnosis of solid pancreaticobiliary lesions: Clinical aspects to improve the diagnosis,” World J. Gastroenterol., vol. 22, no. 2, 2016, Art. no. 628.
[6] S. H. Magee et al., “The presence of a cytopathologist increases the diagnostic accuracy of endoscopic ultrasound-guided fine needle aspiration cytology for pancreatic adenocarcinoma: A meta-analysis,” Cytopathology, vol. 24, no. 3, pp. 159–171, 2013.
[7] H. Matsubayashi et al., “Pathological and molecular aspects to improve endoscopic ultrasonography-guided fine-needle aspiration from solid pancreatic lesions,” Pancreas, vol. 47, no. 2, pp. 163–172, 2018.

[8] T. Hayashi et al., “Rapid on-site evaluation by endosonographer during endoscopic ultrasound-guided fine needle aspiration for pancreatic solid masses,” J. Gastroenterol. Hepatol., vol. 28, no. 4, pp. 656–663, 2013.

[9] J. Iglesias-Garcia et al., “Influence of on-site cytopathology evaluation on the diagnostic accuracy of endoscopic ultrasound-guided fine needle aspiration (EUS-FNA) of solid pancreatic masses,” Official J. Amer. College Gastroenterol., vol. 106, no. 9, pp. 1705–1710, 2011.

[10] F. Yang, E. Liu, and S. Sun, “Rapid on-site evaluation (ROSE) with EUS-FNA: The ROSE looks beautiful,” Endoscopic Ultrasound, vol. 8, no. 5, 2019, Art. no. 283.

[11] P. A. van Riet, D. L. Cahen, J. W. Poley, and M. J. Bruno, “ Mapping international practice patterns in EUS-guided tissue sampling: Outcome of a global survey,” Endoscopy Int. Open, vol. 4, no. 03, pp. E360–E370, 2016.

[12] M. S. Landau and L. Pantanowitz, “Artificial intelligence in cytopathology: A review of the literature and overview of commercial landscape,” J. Amer. Soc. Cytopathol., vol. 8, no. 4, pp. 230–241, 2019.

[13] W. Liu et al., “Is aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers,” Comput. Biol. Med., vol. 141, 2022, Art. no. 105026.

[14] Y. Hashimoto et al., “Mo1296 reliminary result of computer aided diagnosis (CAD) performance using deep learning in EUS-FNA cytology of pancreatic cancer,” Gastrointestinal Endoscopy, vol. 87, no. 6, 2018, Art. no. AB434.

[15] Y. Hashimoto, “Prospective comparison study of EUS-FNA onsite cytopathology diagnosis by pathologist versus our designed deep learning algorithm in suspected pancreatic cancer,” Gastroenterology, vol. 158, no. 6, 2020, Art. no. S-17.

[16] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 2017, 2017, pp. 5999–6009.

[17] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Representations, 2014, pp. 1–14.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2818–2826.

[21] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.

[22] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114.

[23] A. Howard et al., “Searching for mobileNetV3,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 1314–1324.

[24] H. Chen et al., “GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection,” Pattern Recognit., vol. 130, 2022, Art. no. 108827.

[25] A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, “Bottleneck transformers for visual recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 16519–16529.

[26] T. Prange-meier, C. Reich, and H. Koepp, “Attention-based transformers for instance segmentation of cells in microstructures,” in Proc. IEEE Int. Conf. Bioinf. Biomed., 2020, pp. 700–707.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213–229.

[28] L. Yang, R. Zhang, L. Li, and X. Xie, “Simam: A simple, parameter-free attention module for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 11863–11874.

[29] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 8024–8035.

[30] R. Wightman, “Pytorch image models,” 2021. [Online]. Available: https://github.com/rwightman/pytorch-image-models

[31] 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., 2017, pp. 618–626.

[32] S. Woo, J. Park, J. Y. Lee, and J. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

[33] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation net-works,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[34] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 10347–10357.

[35] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10012–10022.

[36] Z. Peng et al., “Conformer: Local features coupling global representations for visual recognition,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 367–376.

[37] C. F. R. Chen, Q. Fan, and R. Panda, “Crosstv: Cross-attention multi-scale vision transformer for image classification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 357–366.

[38] S. Mehta and M. Rastegari, “Mobilevit: Light-weight, general-purpose and mobile-friendly vision transformer,” 2021. [Online]. Available: https://arxiv.org/abs/2110.02178

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.