Cross-Scale Fusion Transformer for Histopathological Image Classification

Sheng-Kai Huang, Yu-Ting Yu, Chun-Rong Huang, *Senior Member, IEEE*, and Hsiu-Chi Cheng

***Abstract*—Histopathological images provide the medical evidences to help the disease diagnosis. However, pathologists are not always available or are overloaded by work. Moreover, the variations of pathological images with respect to different organs, cell sizes and magnification factors lead to the difficulty of developing a general method to solve the histopathological image classification problems. To address these issues, we propose a novel cross-scale fusion (CSF) transformer which consists of the multiple field-of-view patch embedding module, the transformer encoders and the cross-fusion modules. Based on the proposed modules, the CSF transformer can effectively integrate patch embeddings of different field-of-views to learn cross-scale contextual correlations, which represent tissues and cells of different sizes and magnification factors, with less memory usage and computation compared with the state-of-the-art transformers. To verify the generalization ability of the CSF transformer, experiments are performed on four public datasets of different organs and magnification factors. The CSF transformer outperforms the state-of-the-art task specific methods, convolutional neural network-based methods and transformer-based methods.**

***Index Terms*—Deep learning, histopathological image classification, transformer.**

Manuscript received 16 January 2023; revised 3 May 2023, 17 June 2023, and 2 September 2023; accepted 27 September 2023. Date of publication 6 October 2023; date of current version 5 January 2024. This work was supported by the National Science and Technology Council of Taiwan under Grants NSTC 111-2634-F-006-012, NSTC 111-2628-E-006-011-MY3, and NSTC 112-2327-B-006-008. *(Sheng-Kai Huang and Yu-Ting Yu are co-first authors.) (Corresponding author: Chun-Rong Huang.)*

Sheng-Kai Huang is with the Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan (e-mail: g110056163@mail.nchu.edu.tw).

Yu-Ting Yu is with the Department of Pathology, Chung Shan Medical University Hospital, Chung Shan Medical University, Taichung 402, Taiwan (e-mail: yuting.taipei@gmail.com).

Chun-Rong Huang is with the Cross College Elite Program, and Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan 701, Taiwan, and also with Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan (e-mail: crhuang@gs.ncku.edu.tw).

Hsiu-Chi Cheng is with the Department of Internal Medicine, Institute of Clinical Medicine and Molecular Medicine, National Cheng Kung University Hospital, National Cheng Kung University, Tainan 701, Taiwan, and also with the Department of Internal Medicine, Tainan Hospital, Ministry of Health and Welfare, Tainan 701, Taiwan (e-mail: teishuki@mail.ncku.edu.tw).

The source code will be available in our GitHub https://github.com/nchucvml/CSFT.

Digital Object Identifier 10.1109/JBHI.2023.3322387

I. Introduction

HISTOPATHOLOGICAL images have been shown their effectiveness in early diagnosis of various diseases [1]. To identify benign and malignant regions, pathologists need to manually screen the histopathological images. Such screening is very time-consuming and heavily relies on the experience of pathologists. To alleviate the burdens of the pathologists and provide consistent diagnosis results, computer-aided diagnosis (CAD) methods are proposed for histopathological image classification [2], [3], [4], [5], [6], [7].

Among various CAD methods for histopathological image analysis, deep learning-based methods [3], [4], [5], [6], [7], especially the convolutional neural networks (CNNs), have achieved state-of-the-art performance compared with conventional methods. Although CNNs can extract semantic deep features to represent the histopathological images, the field-of-views of the learned features of CNNs are usually limited by the kernel sizes of filters. Moreover, only the features of local regions are extracted. Nevertheless, the sizes of cells and nuclei of tissues are usually variant, and thus these tissues are hard to be represented by CNNs under a fixed field-of-view [8], [9]. Even though deep features can be extracted from parallel CNNs of different field-of-views, these features still only represent local regions without considering context and position information between different regions.

To address the shortcomings of CNN-based methods, vision transformer (ViT) [10] is applied to histopathological image classification [11], [12]. ViT divides the histopathological image into several patches and presents the spatial correlations among patches by using multi-head self-attention. Nevertheless, the learned features of ViT is still under a fixed field-of-view with respect to the patch sizes and thus the scale problem of cells remains. Moreover, ViT requires relatively high computational and memory costs.

Besides the aforementioned issues, the tissues of different datasets are biopsied from different organs and scanned by using different magnification factors. The content variations of different histopathological datasets are also very significant. Thus, many task-specific methods [13], [14] are developed to solve the histopathological classification problems with respect to the datasets of the specific tissues. As shown in [12], conventional transformers are still hard to achieve state-of-the-art performance when they are applied to datasets of different organs and magnification factors. Thus, developing a general histopathological classification method which can be applied to

separately learn different tissue types from different datasets of different magnification factors becomes one of the most challenging issues for histopathological image analysis.

In this article, we propose the cross-scale fusion (CSF) transformer which is a novel transformer structure for general histopathological image classification. The CSF transformer consists of a multiple field-of-view patch embedding module which generates patch embeddings of different field-of-views with respect to histopathological images of different cell sizes and magnification factors. The transformer encoders learn from patch embeddings of different field-of-views at first. The learned patch embeddings are crossly fused by the proposed cross-scale fusion modules to generate cross-scale patch embeddings. In this way, the patch embeddings of the larger field-of-views can help guide the learning of patch embeddings of smaller field-of-views to increase the discriminability of learned features for each transformer encoder.

By introducing the cross-scale fusion module, the cross-scale contextual correlations among patches of different field-of-views can be learned by the transformer encoders. The learned patch embeddings in each layer can be more distinctive to better represent histopathological images of different scales and magnification factors. Based on these distinctive features, we show that the number of transformer encoders can be reduced. Thus, the computational cost and memory usage of the CSF transformer are also reduced compared with the state-of-the-art transformers. Finally, the learned features of different field-of-views are crossly fused again to represent the input histopathological image for classification. To avoid over-fitting and reduce the number of model parameters, we apply the shared-weights scheme to the transformer encoders of the same level. As shown in the experimental results, our method is evaluated on four histopathological datasets of different organs and magnification factors, and outperforms the state-of-the-art task-specific methods, CNN-based methods and transformer-based methods.

Our main contributions are summarized below:

- We propose a novel transformer structure to effectively learn the cross-scale contextual correlations of patches of histopathological images based on different field-of-views generated by the proposed multiple field-of-view patch embedding module in an end-to-end trainable manner.
- Cross-fusion modules integrate patch embeddings of different field-of-views to guide the learning of the patch embeddings of smaller field-of-views by using the patch embeddings of larger field-of-views to generate more robust feature representations. Such design helps reduce the number of model parameters and computation costs compared with state-of-the-art transformers.
- Based on the unique structure of the CSF transformer, we show that it can separately learn different target tissue types and generalize to classify the learned target tissues of various cell sizes and different magnification factors.

In Section II, we review conventional methods and state-of-the-art methods. Section III describes the proposed CSF transformer in details. Experimental results are shown in Section IV. Finally, the conclusions and future work are given in Section V.

II. Related Work

A. Conventional Methods

Conventional methods rely on hand-crafted features [15], [16], [17], [18], [19] for histopathological image classification. To more effectively represent features of histopathological images, sparse dictionary-based methods are proposed. For example, Srinivas et al. [13] propose a simultaneous sparsity model-based multi-channel histopathological image representation and classification (SHIRC) method to represent histopathological images as a sparse linear combination of dictionary atoms composed by training samples. To further enhance class-specific features, a discriminative feature-oriented dictionary learning (DFDL) method [2] and a saliency-based dictionary learning (SDL) method [20] are proposed. Besides the considerations of enhancing different features in dictionary learning, Xiao et al. [21] consider multi-channel joint sparse models to improve the performance. To reduce the computational complexity of sparse dictionary learning, analysis–synthesis model learning with the shared features (ALSF) algorithm is proposed in [22].

Due to the appearance variations of the cells and tissues of the histopathological images, hand-crafted features are hard to effectively represent the appearance variations. Compared with aforementioned conventional methods, deep learning-based methods aim to learn semantic deep features to represent cells and tissues of the histopathological images and thus have been shown more effective for histopathological image classification.

B. Deep Learning Methods

Recently, many deep learning methods especially the convolutional neural networks (CNNs) have been proposed to solve the histopathological image classification problems [14], [23]. To further obtain the global feature distribution of breast cancer histopathological images, task-specific networks are proposed. Song et al. [24] extract fisher vectors from the CNN model (FVCNN) to classify malignant and benign regions of breast histopathological images. Li et al. [25] propose the deep second-order pooling network (DSoPN) based on matrix power normalization which extracts second-order information for global feature representations. To apply hierarchical feature representation of cells, Han et al. [26] propose a class structure-based deep convolutional neural network. Toğaçar et al. [27] propose BreastNet which consists of convolutional, pooling, residual and dense blocks for breast histopathological image classification. Mi et al. [28] adopt a two-stage model to solve the patch-level classification and whole slide image (WSI) classification problems. Zou et al. [29] propose the attention high-order deep network (AHoNet) which adopts ResNet-18 [30] and a channel attention module for breast cancer histopathological image classification. To solve the overfitting problem of CNN, Liu et al. [31] propose the AlexNet-BC model with the penalty of overconfident low-entropy output distributions.

Instead of considering a single network, Yang et al. [32] propose the ensemble of DenseNet [33] and ResNet [30] for breast cancer histopathological image classification. Xu et al. [34] combine a decision network with a soft-attention classification

network. The front network aims to select representative patches based on the latter network for representing breast lesions. Because the computational load of processing the histopathological images, Burçak et al. [35] discuss several optimization schemes to achieve faster learning. To reduce the burden of labeling, Qi et al. [36] propose an entropy-based strategy and a confidence-boosting strategy to update the CNN models with an increasing dataset.

Besides classifying breast cancer lesions from histopathological images, various histopathological classification methods are proposed with respect to different organs. For example, Sun et al. [37] propose a deep learning method by combining transfer learning and multiple-instance learning to obtain patch and image features for liver cancer histopathological classification. However, the end-to-end training is not achieved. Belharbi et al. [4] enable the CNN model to constrain both non-discriminative and discriminative regions simultaneously for colon cancer lesion classification and segmentation. To address the multi-scale changes of cells, Lin et al. [5] propose pyramidal deep-broad learning to obtain the pyramidal contextual information based on a multi-resolution image pyramid for lung and colon histopathological image classification.

More recently, transformer-based histopathological image classification methods are proposed. Shao et al. [38] propose a multiple instance learning method based on the transformer to explore correlations between different patches of histopathological images. Chen et al. [39] propose the hierarchical image pyramid transformer (HIPT) based on vision transformer [10] by imposing the hierarchical structure in WSIs. In [11], Zou et al. combine transformer and CNN to develop a dual-stream convolution expanded transformer network (DCET-Net) for histopathological image classification. Because transformer can capture the relationship of different patches in the WSI, a transformer-guided framework [40] is proposed to achieve attention-based mutual knowledge distillation for diagnosing lymph node metastasis. In [12], the performance of vision transformer [10] and Swin transformer [41] are compared with respect to different histopathological image datasets. Tavolara et al. [42] develop an automated method based on Swin transformer [41] to identify tumor budding and generate ground truth by using H&E stained slides.

Compared with aforementioned deep learning methods which extract local features to represent tissues, the proposed method not only extracts spatial contextual information between different histopathological regions via the multi-head self-attention scheme, but also applies the cross-scale contextual correlations to guide the feature learning based on the patch embeddings of different field-of-views. As a result, the proposed method can successfully solve the histopathological image classification problems of different tissues and magnification factors.

III. Method

A. Overview

An overview of the proposed cross-scale fusion transformer is shown in Fig. 1. Given an input image I of size $H \times W \times C$, where H and W denote the image height and width, and C is the number of channels of I. I serves as the input of the multiple field-of-view patch embedding module to obtain the patching embedding $\boldsymbol{z}_0^i$, where $i \in \{s, m, l\}$, and s, m and l represent the patch embeddings of small, medium and large field-of-views, respectively. The CSF transformer contains 9 transformer encoders as shown in Fig. 1. Each transformer encoder $\mathcal{T}_\ell^i$ aims to learn the patch embedding $\boldsymbol{z}_\ell^i$ in the level ℓ with respect to the field-of-view i from the patch embedding generated in the level $\ell - 1$. To avoid over-fitting and reduce the number of learned parameters, the transformer encoders of the same level ℓ have shared-weights.

With the learned transformer encoder features of different field-of-views, i.e. small, medium and large field-of-views, the cross-fusion modules are proposed to crossly integrate these features based on the levels and field-of-views of the network. The first kind of the cross-fusion module $\mathcal{F}_\ell^{sm}$ aims to crossly fuse the features of the small and medium field-of-views in the level ℓ. The second kind of the cross-fusion module $\mathcal{F}_\ell^{ml}$ aims to crossly fuse the features of the medium and large field-of-views in the level ℓ. Finally, to integrate features of all field-of-views, the third kind of the cross-fusion module $\mathcal{F}_\ell^{sml}$ is proposed. These cross-fusion modules generate cross-scale patch embeddings to help the transformer encoders learn the cross-scale contextual correlations among patches of different field-of-views and retrieve the most representative features of I under different field-of-views and magnification factors for histopathological image classification. Finally, the last cross-scale patch embedding is passed to a channel-wise average pooling layer and a fully-connected layer to compute the weighted cross entropy loss $\mathcal{L}_{wce}$ to drive the update of the network.

B. Cross-Scale Fusion Transformer

Conventionally, the resolution of the input patches of the transformer encoder [10] is fixed which implies that the field-of-view of the learned features is also fixed. In practice, the sizes of cells of histopathological images are variant and are also affected by the magnification factors. Learning from histopathological images based on the unified field-of-view may easily fail when classifying cells of different sizes and magnification factors. A naive method to generate multi-scale features is to use parallel transformers where each transformer contains the same number of encoders. However, given the high computational complexity of parallel transformers and multi-head self-attentions, this brute-force method will lead to an explosion of the memory usage, number of parameters and computational cost. To solve the problems, we propose the CSF transformer which can efficiently and effectively learn cross-scale contextual correlations among transformer encoders of different field-of-views to reduce the computational complexity and the number of transformer encoders compared with the naive parallel transformers. In the following, we will introduce each module in the CSF transformer.

1) Multiple Field-of-View Patch Embedding Module: The performance of the standard transformer [10] is limited to patches with a fixed field-of-view. In order to effectively learn feature representations of different field-of-views, we propose

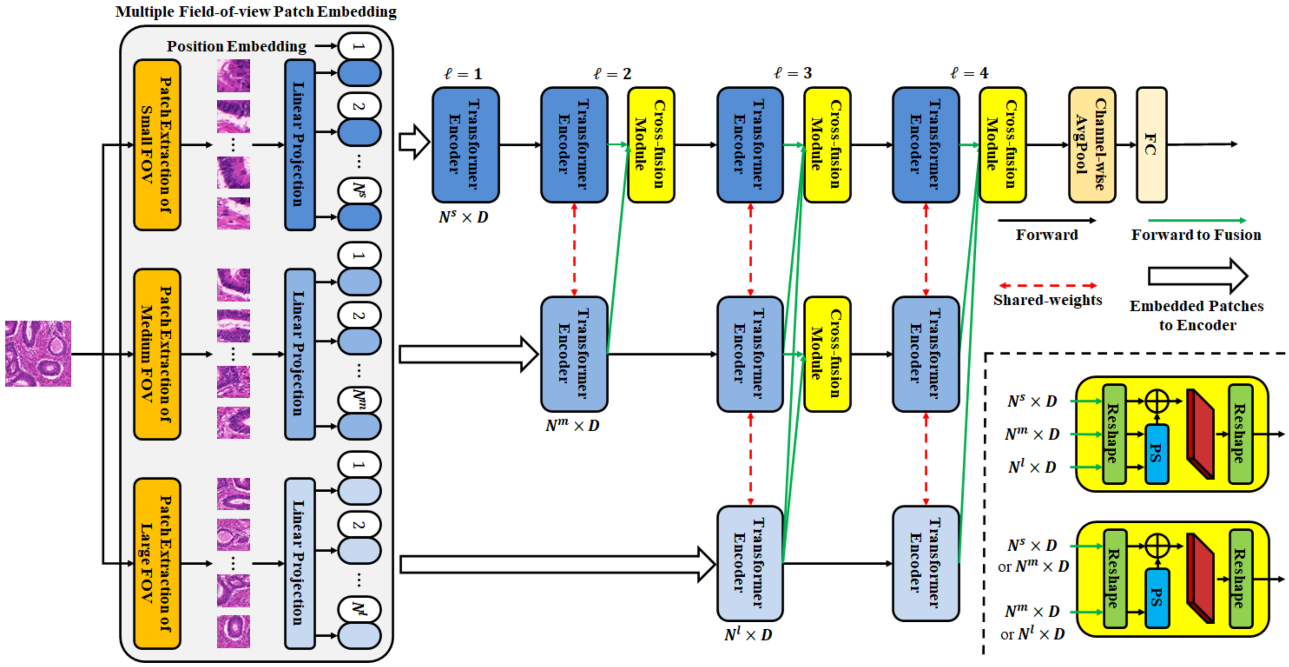

Fig. 1. Overview of the proposed cross-scale fusion (CSF) transformer. The CSF transformer consists of a multiple field-of-view (FOV) patch embedding module which generates patch embeddings of different field-of-views for transformer encoders. Each transformer encoder aims to learn features from different patch embeddings generated by the transformers or the cross-scale fusion modules in the previous level via multi-head self-attention. The cross-scale fusion modules aim to generate cross-scale patch embeddings for the transformer encoders of the next level to increase the discriminability of learned features by crossly fusing features of different field-of-views. The detailed structures of the cross-scale fusion modules are shown in the bottom right region of the figure. The pixel shuffle (PS) composes new feature maps of the same spatial dimension by using sub-pixel convolution. Finally, the learned features of different field-of-views are crossly fused again with a channel-wise average pooling layer and a fully connected layer for prediction.

multiple field-of-view patch embedding module as shown in Fig. 1. Because the resolution of the input image I will be different for each histopathological dataset, we resize I into $M \times M$ at first, where M is 256. In the following, we will introduce how to obtain patch embeddings of different field-of-views.

Let P be the patch size and the number N^i of the patches with respect to the field-of-view i can be computed as $N^i = (\frac{M}{\lambda^i \cdot P})^2$, where $\lambda^i = 1, 2$ and 4 for the small, medium and large field-of-views, respectively. Then, we flatten each patch to $\boldsymbol{x}_n^i \in \mathbb{R}^{1\times((\lambda^i \cdot P)^2 \cdot C)}$, where $\boldsymbol{x}_n^i$ is a 1-D vector of the nth patch with respect to the field-of-view i. To obtain the patch embedding for the following transformer encoders, a learnable linear projection $\boldsymbol{e}^i$ is applied to each $\boldsymbol{x}_n^i$. A learnable position embedding $\boldsymbol{e}_{pos}^i$ as shown in vision transformer [10] is also appended with the patch embedding to indicate the positional information with respect to the field-of-view i. Appending the position embedding to each patch can help maintain the position information. The output embedding $\boldsymbol{z}_0^i$ for I with respect to the field-of-view i is defined as follows:

$$\boldsymbol{z}_0^i = [\boldsymbol{x}_1^i \boldsymbol{e}^i; \boldsymbol{x}_2^i \boldsymbol{e}^i; \ldots; \boldsymbol{x}_{N^i}^i \boldsymbol{e}^i] + \boldsymbol{e}_{pos}^i, \tag{1}$$

where $\boldsymbol{e}^i \in \mathbb{R}^{((\lambda^i \cdot P)^2 \cdot C)\times D}$ is the learnable linear projection and $\boldsymbol{e}_{pos}^i \in \mathbb{R}^{N^i \times D}$ is the position embedding with respect to the field-of-view i, and $D = 768$. In this way, the multiple field-of-view patch embedding module generates the patch embedding $\boldsymbol{z}_0^i$ with respect to the field-of-view i.

Please note that because the patch embeddings of different field-of-views will be fused to extract more representative features in the following, the class token is omitted to maintain the dimension consistency for embeddings of different field-of-views.

2) Transformer Encoder: Each transformer encoder $\mathcal{T}_\ell^i$ aims to learn representative features by using multi-head self-attention. Let $\boldsymbol{z}_{\ell-1}^i$ be the input patch embedding generated by the transformer encoder or the cross-fusion module with respect to the field-of-view i in the level $\ell - 1$. Each transformer encoder contains a normalization layer $LN(\cdot)$ followed by a multi-head self-attention layer $MSA(\cdot)$ to compute the multi-head self-attention. A residual connection of the input is combined with the output of the multi-head self-attention layer to generate $\boldsymbol{z'}_\ell^i$ as follows:

$$\boldsymbol{z'}_\ell^i = MSA(LN(\boldsymbol{z}_{\ell-1}^i)) + \boldsymbol{z}_{\ell-1}^i. \tag{2}$$

Here $MSA(\cdot)$ is defined as:

$$MSA(\boldsymbol{Q}_\ell^i, \boldsymbol{K}_\ell^i, \boldsymbol{V}_\ell^i) = softmax\left(\frac{\boldsymbol{Q}_\ell^i {\boldsymbol{K}_\ell^i}^T}{\sqrt{\boldsymbol{d}}}\right)\boldsymbol{V}_\ell^i. \tag{3}$$

Please note that the output of $LN(\boldsymbol{z}_{\ell-1}^i)$ is multiplied by three different learnable projection matrices $\boldsymbol{W}_{Q,\ell}^i$, $\boldsymbol{W}_{K,\ell}^i$ and $\boldsymbol{W}_{V,\ell}^i$ to obtain $\boldsymbol{Q}_\ell^i$, $\boldsymbol{K}_\ell^i$, and $\boldsymbol{V}_\ell^i$, respectively to perform the multi-head self-attention operation. We employ the number of head $n_{head} = 12$ and $\boldsymbol{d} = D/n_{head} = 64$. $\boldsymbol{z'}_\ell^i$ is passed to the second

normalization layer and a feed forward layer with a residual connection to obtain the output of the transformer encoder as follows:

$$\boldsymbol{z}_\ell^i = FF(LN(\boldsymbol{z}'^i_\ell)) + \boldsymbol{z}'^i_\ell, \tag{4}$$

where $FF(\cdot)$ represents the feed forward layer.

When the field-of-view is small, more patches are obtained. To better represent the spatial correlations of these patches by using the CSF transformer, we apply four transformer encoders to learn features from the small field-of-view. In contrast, the numbers of patches of the medium and large field-of-views are fewer. Thus, we apply three and two transformer encoders to learn features from the medium and large field-of-views, respectively. By this design, we can reduce the number of transformer encoders compared with conventional transformers, and also benefit the memory usage and computational efficiency.

3) Cross-Fusion Module: To further discover cross-scale contextual correlations between different regions of histopathological images with respect to different field-of-views by using the transformer encoders, the cross-fusion modules are proposed for the cross-scale patch embedding generation. In the module, we integrate the features of the larger field-of-views with the features of smaller field-of-view to assist the feature learning of the transformer encoders of the smaller field-of-view. In this way, the larger field-of-view information can guide the learning of the smaller field-of-view and further help discover the cross-scale contextual correlations from cross-scale patch embeddings. In the following, we will describe three kinds of cross-fusion modules to fuse patch embeddings of different field-of-views.

The first kind of the cross-fusion module $\mathcal{F}_\ell^{sm}$ aims to fuse two patch embeddings $\boldsymbol{z}_\ell^s$ and $\boldsymbol{z}_\ell^m$ of the small and medium field-of-views in the level ℓ to generate cross-scale patch embedding to assist the feature learning of the transformer encoder $\mathcal{T}_{\ell+1}^s$ of the small field-of-view in the level $\ell+1$. To ensure that the cross-scale contextual correlations of cross-scale patch embedding can be learned by the transformer encoder, we need to correctly remap and integrate patch embeddings of different field-of-views. To achieve the goal, we first reshape $\boldsymbol{z}_\ell^s \in \mathbb{R}^{N^s \times D}$ and $\boldsymbol{z}_\ell^m \in \mathbb{R}^{N^m \times D}$ from 1-D feature vectors to 2-D feature maps $R(\boldsymbol{z}_\ell^s) \in \mathbb{R}^{\sqrt{N^s}\times\sqrt{N^s}\times D}$ and $R(\boldsymbol{z}_\ell^m) \in \mathbb{R}^{\sqrt{N^m}\times\sqrt{N^m}\times D}$, where $R(\cdot)$ is the 2-D reshape function, $N^s = (\frac{M}{P})^2$ and $N^m = (\frac{M}{2P})^2$. In this way, $R(\boldsymbol{z}_\ell^s)$ and $R(\boldsymbol{z}_\ell^m)$ can be considered as the spatial feature maps to represent features of the small and medium field-of-views, respectively. However, because $\boldsymbol{z}_\ell^s$ and $\boldsymbol{z}_\ell^m$ are learned from patches of different field-of-views, their numbers of patches are different, i.e. the dimensions of $R(\boldsymbol{z}_\ell^s)$ and $R(\boldsymbol{z}_\ell^m)$ are different. Thus, $R(\boldsymbol{z}_\ell^s)$ and $R(\boldsymbol{z}_\ell^m)$ cannot be directly fused.

While naively upsampling of $R(\boldsymbol{z}_\ell^m)$ leads to blur effects, we consider using pixel shuffle [43] for feature remapping of $R(\boldsymbol{z}_\ell^m)$. Pixel shuffle is a periodic shuffling operator that rearranges the elements of $R(\boldsymbol{z}_\ell^m)$ to compose new feature maps which have the same spatial dimension as $R(\boldsymbol{z}_\ell^s)$ by using sub-pixel convolution [43]. Pixel shuffle will change the dimension $\mathbb{R}^{\sqrt{N^m}\times\sqrt{N^m}\times D}$ of $R(\boldsymbol{z}_\ell^m)$ to $\mathbb{R}^{\sqrt{N^s}\times\sqrt{N^s}\times D'}$ of $PS(R(\boldsymbol{z}_\ell^m))$, where $PS(\cdot)$ is the pixel shuffle function. In other words, pixel shuffle rearranges the elements of a tensor of $\mathbb{R}^{\sqrt{N^m}\times\sqrt{N^m}\times D}$ to a new tensor of $\mathbb{R}^{\sqrt{N^s}\times\sqrt{N^s}\times D'}$, which is equivalent to $\mathbb{R}^{\sqrt{N^m\theta}\times\sqrt{N^m\theta}\times(D/\theta)}$, where $\theta = \frac{N^s}{N^m}$ is the ratio between the spatial dimension of $R(\boldsymbol{z}_\ell^s)$ and $R(\boldsymbol{z}_\ell^m)$. Because of the design of the multiple field-of-view patch embedding module, the pixel shuffle can be achieved via the sub-pixel convolution in the $R(\boldsymbol{z}_\ell^m)$ space and produces high resolution 2-D feature maps from $R(\boldsymbol{z}_\ell^m)$ directly with one upscaling filter. After pixel shuffle, the channel features of the same patches are interpolated to generate high resolution feature maps to represent $\boldsymbol{z}_\ell^m$.

To integrate the patch embeddings of different field-of-views, we concatenate $PS(R(\boldsymbol{z}_\ell^m))$ to $R(\boldsymbol{z}_\ell^s)$ and obtain $\boldsymbol{f}_\ell^{sm}$ as follows:

$$\boldsymbol{f}_\ell^{sm} = R(\boldsymbol{z}_\ell^s) \oplus PS(R(\boldsymbol{z}_\ell^m)), \tag{5}$$

where $\oplus$ represents the feature concatenation operator, and the dimension of $\boldsymbol{f}_\ell^{sm} \in \mathbb{R}^{\sqrt{N^s}\times\sqrt{N^s}\times(D+D/\theta)}$. Pixel shuffle maintains the positions of patches, so the fused feature maps will also maintain the spatial correlations. Thus, the following transformer encoder can then learn the cross-scale contextual correlations based on the fused feature maps.

Because pixel shuffle only provides upsampled results based on the reshaped feature maps, the correlations among different channels of $\boldsymbol{f}_\ell^{sm}$ are not well discovered. Moreover, the number of channels increases which will also lead to the increasing memory usage of the following transformer encoder. To address these issues, a learnable 3×3 convolutional layer with padding 1 is then used to learn the new feature maps of D channels from $\boldsymbol{f}_\ell^{sm}$. The convoluted feature maps are reshaped to a 1-D vector to obtain the cross-scale patch embedding $\boldsymbol{z}_\ell^{sm}$ which serves as the input of the transformer encoder $\mathcal{T}_{\ell+1}^s$. $\boldsymbol{z}_\ell^{sm}$ is defined as follows:

$$\boldsymbol{z}_\ell^{sm} = R'(conv^3(\boldsymbol{f}_\ell^{sm})), \tag{6}$$

where $R'(\cdot)$ is the 1-D reshape function, $conv^3(\cdot)$ is a learnable 3×3 convolutional layer, and the dimension of $\boldsymbol{z}_\ell^{sm}$ is $\mathbb{R}^{N^s \times D}$.

The second kind of the cross-fusion module $\mathcal{F}_\ell^{ml}$ aims to fuse two patch embeddings $\boldsymbol{z}_\ell^m$ and $\boldsymbol{z}_\ell^l$ of the medium and large field-of-views in the level ℓ. Similar to $\mathcal{F}_\ell^{sm}$, we concatenate $PS(R(\boldsymbol{z}_\ell^l))$ to $R(\boldsymbol{z}_\ell^m)$ to assist the learning of the transformer encoder $\mathcal{T}_{\ell+1}^m$ and obtain $\boldsymbol{f}_\ell^{ml}$ as follows:

$$\boldsymbol{f}_\ell^{ml} = R(\boldsymbol{z}_\ell^m) \oplus PS(R(\boldsymbol{z}_\ell^l)), \tag{7}$$

where the dimension of $\boldsymbol{f}_\ell^{ml} \in \mathbb{R}^{\sqrt{N^m}\times\sqrt{N^m}\times(D+D/\theta')}$, $\theta' = \frac{N^m}{N^l}$ and $N^l = (\frac{M}{4P})^2$. The cross-scale patch embedding $\boldsymbol{z}_\ell^{ml}$ for the next transformer encoder $\mathcal{T}_{\ell+1}^m$ is then defined as follows:

$$\boldsymbol{z}_\ell^{ml} = R'(conv^3(\boldsymbol{f}_\ell^{ml})), \tag{8}$$

where the dimension of $\boldsymbol{z}_\ell^{ml}$ is $\mathbb{R}^{N^m \times D}$.

The third kind of the cross-fusion module $\mathcal{F}_\ell^{sml}$ aims to fuse three patch embeddings $\boldsymbol{z}_\ell^s$, $\boldsymbol{z}_\ell^m$ and $\boldsymbol{z}_\ell^l$ of the small, medium and large field-of-views in the level ℓ. Similar to the process mentioned above, the patch embeddings of the medium and large field-of-views are reshaped and then resized by using pixel shuffle functions. Then, we concatenate $PS(R(\boldsymbol{z}_\ell^m))$ and

$PS(R(\boldsymbol{z}_\ell^l))$ to $R(\boldsymbol{z}_\ell^s)$ and obtain $\boldsymbol{f}_\ell^{sml}$ as follows:

$$\boldsymbol{f}_\ell^{sml} = R(\boldsymbol{z}_\ell^s) \oplus PS(R(\boldsymbol{z}_\ell^m) \oplus PS(R(\boldsymbol{z}_\ell^l)), \tag{9}$$

where the dimension of $\boldsymbol{f}_\ell^{sml}$ is $\mathbb{R}^{\sqrt{N^s}\times\sqrt{N^s}\times(D+D/\theta+D/\theta'')}$, $\theta'' = \frac{N^s}{N^l}$. $\boldsymbol{f}_\ell^{sml}$ is also passed to a learnable 3×3 convolutional layer with padding 1 and D channels, and then is reshaped to obtain the cross-scale patch embedding $\boldsymbol{z}_\ell^{sml}$ for the input of $\mathcal{T}_{\ell+1}^s$ in the next level as follows:

$$\boldsymbol{z}_\ell^{sml} = R'(conv^3(\boldsymbol{f}_\ell^{sml})), \tag{10}$$

where the dimension of $\boldsymbol{z}_\ell^{sml}$ is also $\in \mathbb{R}^{N^s\times D}$ to achieve the dimension consistency of patch embeddings.

The last cross-fusion patch embedding is passed to a channel-wise average pooling layer and a fully connected layer to compute the loss and obtain the prediction results. Because the distributions of cells of the the histopathological images may be imbalanced, we apply weighted cross-entropy loss function $\mathcal{L}_{wce}(\cdot)$ which is defined as follows:

$$\mathcal{L}_{wce} = -\sum_{c=1}^{C} \frac{N}{N_c}(y_c \log(\hat{y}_c)) \tag{11}$$

where C is the number of classes, N is the number of training images, N_c is number of training images of class c, y_c is the ground truth label of the input image I and $\hat{y}_c$ is the prediction of the network.

C. Implementation Details

The input image is resized to 256×256 and the patch size P is 16. Our transformer encoder is pre-trained on ImageNet [44]. All the datasets are trained by using SGD with the momentum 0.9 and a weight decay 5×10^{-4} with the weighted cross-entropy loss. The batch size is 285. Our method is implemented by using PyTorch 1.12 and is run under a server with a NVIDIA Tesla V100 GPU. Because the BreakHis dataset is relatively large, the learning rate is set to 10^{-3} and the number of epochs is set to 100. The learning rate of the remaining dataset is set to 10^{-5} and the number of epochs is set to 1000. To improve the generalization ability of the learned features, we also apply data augmentations including random cropping, horizontal flipping, vertical flipping, and rotations of 90, 180 and 270 degrees during training.

IV. Experimental Results

A. Dataset

Our goal is to evaluate the generalization ability of the proposed method by using different histopathological image classification datasets of different tissues. Therefore, four public histopathological datasets, BreakHis [45], ADL [13], YTMF [46] and GlaS [47], [48], with detailed image descriptions were employed for evaluations. Fig. 2 shows sample histopathological images of these datasets. Please note that each dataset has its own training and testing data partition criteria, and evaluation metrics due to the divergence of different tissues and data providers. Thus, we followed individual dataset settings provided by each dataset for fair comparisons.

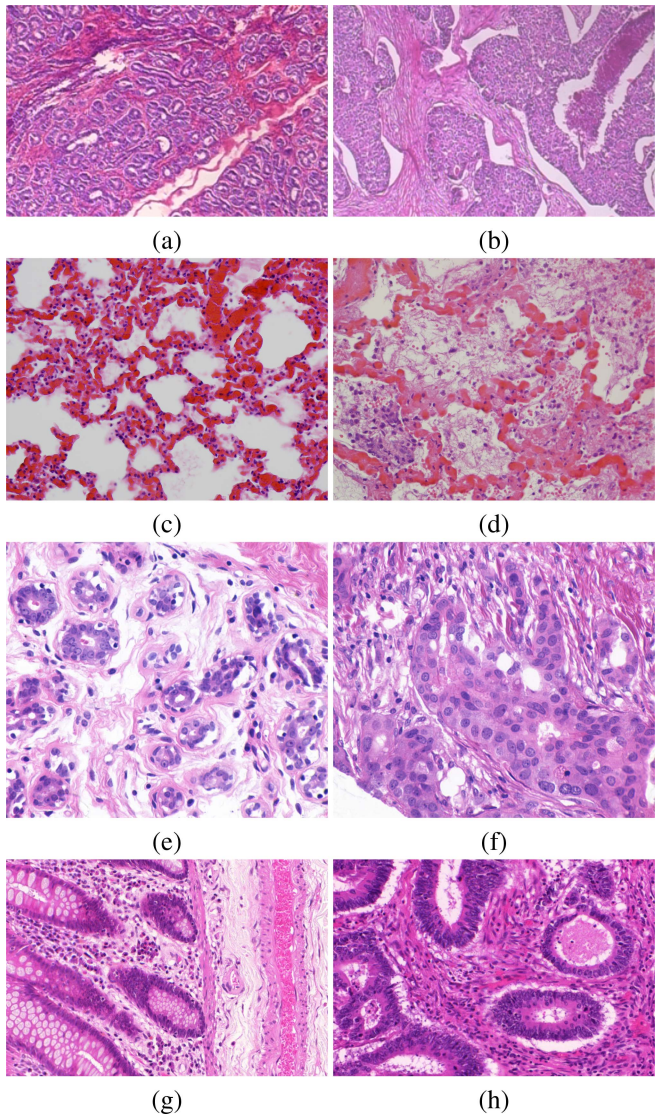

Fig. 2. Sample histopathological images. (a) A benign image of the BreakHis dataset, (b) a malignant image of the BreakHis dataset, (c) a normal image of the lung of the ADL dataset, (d) an inflammation image of the lung of the ADL dataset, (e) a benign image of the YTMF dataset, (f) a malignant image of the YTMF dataset, (g) a benign image of the GlaS dataset and (h) a malignant image of the GlaS dataset.

1) BreakHis dataset [45]: The breast cancer histopathological image classification (BreakHis) dataset was collected from 82 patients using different magnification factors including 40×, 100×, 200× and 400×. The dataset contains 5429 malignant images of tumor tissues and 2480 benign images of normal tissues. For fair comparisons, we adopt the same experimental settings in [25], [45] which randomly divide the dataset into the training set and testing set by 70% and 30%.

In the BreakHis dataset, there are two evaluation metrics. The first one is the image level recognition rate I_{rr} [45], which does not consider patient information. Let N_I be the number

of images of the testing set and N_{rec} is the number of correctly classified images and then I_{rr} can be defined as follows:

$$I_{rr} = \frac{N_{rec}}{N_I}. \tag{12}$$

Because the decision of the breast cancer is patient-wise [45], the second metric is the patient-level recognition rate P_{rr}, which provides the average classification accuracy of all patients based on the classification accuracy of images of each patient. Let N^k_{rec} be the number of correctly classified images and N^k_P be the total number of images of the kth patient. The patient score P^k of the kth patient is defined as:

$$P^k = \frac{N^k_{rec}}{N^k_P}. \tag{13}$$

Then, the patient recognition rate P_{rr} is defined as:

$$P_{rr} = \frac{\sum P^k}{N_P}, \tag{14}$$

where N_P is the total number of patients.

2) ADL dataset [13]: This dataset was provided by pathologists at the animal diagnostics lab (ADL), Pennsylvania State University. These Hematoxylin and Eosin (H&E) stained tissue images were obtained from three different bovine organs including the kidney, lung and spleen. Images of each organ are divided into two categories containing normal and inflammation tissues. Each tissue was scanned by using a whole slide digital scanner with the 40× optical magnification factor. Each class contains approximately 150 images, 40 normal tissue images and 40 inflammation tissue images were used for training and the remaining images were used for testing. Based on the experimental settings in [13], the confusion matrices of each organ are reported.

3) YTMF dataset [46]: The dataset was obtained from the Yale tissue microarray facility (YTMF). This dataset contains 58 Hematoxylin and Eosin (H&E) stained histopathological images of breast cancer cells. The dataset contains 26 malignant images of cancer cells and 32 benign images of normal cells. Because the number of images in this dataset is relatively small, the state-of-the-art methods [20], [22] performed leave-one-out evaluation and showed the confusion matrices for both categories.

4) GlaS dataset [47], [48]: Besides aforementioned tissues of breast, kidney, lung and spleen, we further evaluated the proposed method on colon tissues. The GlaS dataset contains 165 images extracted from Hematoxylin and Eosin (H&E) stained colon histology tissues. The slides were scanned at the 20× magnification factor. The dataset contains 91 malignant images of colon tumor tissues and 74 benign images of normal tissues. Based on the experimental settings in [48], the training set contains 37 benign images and 48 malignant images. The remaining images are used for testing and the total accuracy is reported.

B. Quantitative Results

In the experiments, the proposed method is compared with the state-of-the-art methods for each dataset. For fair comparisons, we followed the specific evaluation metrics in each dataset. Please note that the compared methods including task-specific classification methods which were only able to be applied to the specific dataset. Thus, the compared methods among different datasets may be different. In addition, the reported metrics of the competing methods are retrieved from the original papers. We also run vision transformer (ViT) [10], Swin transformer [41] and HIPT [39] under the same dataset settings, augmentations and hardware for fair comparisons.

Table I shows the compared results including Spanhol et al. [45], Spanhol et al. [23], FVCNN [24], Han et al. [26], Hou [14], Toğaçar et al. [27], Li et al. [25], Mi et al. [28], Zou et al. [29], Burçak et al. [35], AlexNet-BC [31], vision transformer (ViT) [10], Swin transformer [41], DCET-Net [11], HIPT [39] and the proposed method for the BreakHis dataset. This dataset contains histopathological images of four different magnification factors of the breast tissues. Some state-of-the-art methods do not provide the accuracy with respect to patients, so we leave "-" to the metrics of these methods. Because the proposed CSF transformer can simultaneously consider the representation of different field-of-views by using cross-scale scheme, it achieves good performance for classifying histopathological images at different magnification factors. As a result, the proposed method achieves the best classification results with respect to both images and patients as shown in Table I even compared with the task-specific methods.

In addition, we also compared the state-of-the-art transformers including vision transformer (ViT) [10], Swin transformer [41], DCET-Net [11], and HIPT [39]. ViT and Swin transformer did not consider patches of different field-of-views and thus cannot integrate features of different field-of-views to represent cells of different sizes. Although they perform better results compared with recent methods, the proposed method still outperforms them. HIPT [39] considers a hierarchical structure to obtain representations of different levels including cell-level, patch-level and region-level. However, such hierarchical representations may not exist under different magnification factors and thus HIPT achieves worse results. By considering the information of CNNs and transformer, DCET-Net [11] achieves better results. Nevertheless, the proposed method still outperforms these transformer-based methods.

The ADL dataset contains three different kinds of pathological tissues of kidney, lung and spleen for inflammation classification. The confusion matrices of the state-of-the-art methods and the proposed method with respect to each organ are shown in Table II. The state-of-the-art methods including WND-CHARM [16], SHIRC [13], SCD [49], SLIDE [50], DFDL [2], SDL [20], ALSF [22], SDPSD + MIMCJSM [21], FVCNN [24], DEFNet [51], ViT [10], Swin transformer [41], and HIPT [39] are compared. While most methods including the proposed method apply 80 images (40 normal tissue images and 40 inflammation tissue images) for training, many competing methods apply 180 images for training which help their methods to learn well due to more training images. We mark '*' for their unique settings. As shown in Table II, the transformer-based methods, ViT and Swin transformer, achieve better results compared with previous methods because the multi-head self-attention schemes of the transformers help learn the spatial correlations to

TABLE I
COMPARISONS OF THE BREAKHIS DATASET BY USING RECOGNITION RATES (%)

Method	Image Level				Patient Level			
	40×	100×	200×	400×	40×	100×	200×	400×
Spanhol et al. [45]	-	-	-	-	81.60	79.90	85.10	82.30
Spanhol et al. [23]	85.60	83.50	82.70	80.70	90.00	88.40	84.60	86.10
FVCNN [24]	87.00	86.20	85.20	82.90	90.00	88.90	86.90	86.30
Han et al. [26]	95.80	96.90	96.70	94.90	97.10	95.70	96.50	95.70
Hou [14]	90.89	90.99	91.00	90.97	91.00	91.00	91.00	91.00
Toğaçar et al. [27]	97.99	97.84	98.51	95.88	-	-	-	-
Li et al. [25]	96.00	96.16	98.01	95.97	95.01	96.84	97.92	96.28
Mi et al. [28]	96.70	97.60	95.00	93.30	-	-	-	-
Zou et al. [29]	97.58	97.47	99.08	96.52	96.79	97.41	99.29	97.16
Burçak et al. [35]	97.00	97.00	96.00	96.00	-	-	-	-
AlexNet-BC [31]	98.15	97.71	97.96	98.48	-	-	-	-
ViT [10]	97.94	96.86	98.71	96.82	97.69	96.77	99.15	95.63
Swin [41]	97.42	98.68	98.71	98.59	98.18	98.69	98.21	98.68
DCET-Net [11]	**99.00**	98.08	99.34	98.72	98.88	97.94	99.23	99.03
HIPT [39]	80.41	84.13	87.74	85.87	84.09	84.27	87.12	85.98
Proposed	98.11	**99.01**	**99.52**	**99.47**	**98.89**	**99.22**	**99.58**	**99.46**

The bold entities indicate the best results.

TABLE II
COMPARISONS OF THE ADL DATASET BY USING THE CONFUSION MATRIX IN (%)

Method		Kidney		Lung		Spleen	
	Class	Inflammation	Normal	Inflammation	Normal	Inflammation	Normal
WND-CHARM [16]	Inflammation	-	-	-	-	88.2	11.8
SHIRC [13]		83.3	16.7	85.0	15.0	88.3	11.7
SCD* [49]		82.3	17.7	41.0	59.0	65.0	35.0
SLIDE* [50]		24.5	75.5	92.3	7.7	57.5	42.5
DFDL [2]		90.0	10.0	97.4	2.6	92.0	8.0
SDL* [20]		97.8	2.2	97.4	2.6	95.0	5.0
ALSF [22]		-	-	-	-	96.6	3.4
SDPSD + MIMCJSM* [21]		90.3	9.7	98.9	1.1	93.2*	6.2*
FVCNN [24]		-	-	-	-	94.1	6.9
DEFNet [51]		98.6	1.4	96.5	3.5	98.3	1.7
ViT [10]		92.0	8.0	**100.0**	**0.0**	97.5	2.5
Swin [41]		98.6	1.4	97.4	2.6	98.3	1.7
HIPT [39]		77.5	22.5	92.2	7.8	98.3	1.7
Proposed (Cases)		**99.3** (137/138)	**0.7** (1/138)	**100.0** (115/115)	**0.0** (0/115)	**99.2** (118/119)	**0.8** (1/119)
WND-CHARM [16]	Normal	-	-	-	-	19.8	80.2
SHIRC [13]		17.5	82.5	25.0	75.0	35.0	65.0
SCD* [49]		73.0	27.0	35.9	64.1	60.8	39.2
SLIDE* [50]		15.0	85.0	67.3	32.7	56.1	43.9
DFDL [2]		11.8	88.2	3.5	96.5	7.1	92.9
SDL* [20]		12.5	87.5	2.6	97.4	9.8	90.2
ALSF [22]		-	-	-	-	6.6	93.4
SDPSD + MIMCJSM* [21]		9.9	90.1	1.0	99.0	5.8	94.2
FVCNN [24]		-	-	-	-	10.9	90.1
DEFNet [51]		3.4	96.6	1.8	98.2	2.5	97.5
ViT [10]		9.4	90.6	11.5	88.5	6.6	93.4
Swin [41]		**1.7**	**98.3**	0.9	99.1	**1.6**	**98.4**
HIPT [39]		11.1	88.9	16.8	83.2	12.4	87.6
Proposed (Cases)		3.4 (4/117)	96.6 (113/117)	**0.0** (0/113)	**100.0** (113/113)	**1.6** (2/121)	**98.4** (119/121)

represent these histopathological cells. Nevertheless, cross-scale contextual correlations is not considered in ViT, Swin transformer, and HIPT [39]. Thus, the proposed method achieves the best results in most cases of the ADL dataset.

Table III shows the confusion matrices of different methods including WND-CHARM [16], DFDL [2], SDL [20], ALSF [22], FVCNN [24], ViT [10], Swin transformer [41], HIPT [39] and the proposed method. Because more normal training images are available in the YTMF dataset, previous methods performed worse results on the malignant category of the dataset. Again, based on the cross-scale contextual correlations, the proposed method achieves the best results.

Table IV shows the comparison results of different methods of the GlaS dataset. We compared the proposed method with ERASE [52], Wildcat [53], Deep MIL [54], DEFNet [51], EEM [4], ViT [10], Swin transformer [41], and HIPT [39]. The results in Table IV show the performance of the task-specific method EEM for histopathological classification in the GlaS dataset. In comparison, ViT, Swin transformer, and HIPT [39] achieve worse results. Nevertheless, the proposed method can still achieve the best accuracy, which indicates the effectiveness of the proposed transformer structure compared with the remaining transformer-based methods for histopathological image classification.

TABLE III
COMPARISONS OF THE YTMF DATASET BY USING THE CONFUSION MATRIX IN (%)

Method	Class	Malignant	Benign
WND-CHARM [16]	Malignant	46.15	53.85
DFDL [2]		61.54	38.46
SDL [20]		61.54	38.46
ALSF [22]		65.38	34.62
FVCNN [24]		65.38	34.62
ViT [10]		61.54	38.46
Swin [41]		65.38	34.62
HIPT [39]		69.23	30.77
Proposed		**80.77**	**19.23**
(Cases)		(21/26)	(5/26)
WND-CHARM [16]	Benign	21.88	78.12
DFDL [2]		15.62	84.38
SDL [20]		9.38	90.62
ALSF [22]		9.38	90.62
FVCNN [24]		12.50	87.50
ViT [10]		9.38	90.62
Swin [41]		9.38	90.62
HIPT [39]		18.75	81.25
Proposed		**6.25**	**93.75**
(Cases)		(2/32)	(30/32)

TABLE IV
COMPARISONS OF THE GLAS DATASET BY USING ACCURACY (%)

Method	Accuracy
ERASE [52]	92.50
Wildcat [53]	98.75
Deep MIL [54]	97.50
DEFNet [51]	95.00
EEM [4]	**100.00**
ViT [10]	92.50
Swin [41]	97.50
HIPT [39]	67.50
Proposed	**100.00**
(Cases)	(80/80)

TABLE V
PARAMETERS AND FLOPS

Method	Image size	Params	FLOPs
ViT [10]	224×224	86M	17.5G
ViT [10]	384×384	86M	55.4G
Swin [41]	224×224	88M	15.4G
Swin [41]	384×384	88M	47.0G
Proposed	256×256	**57M**	**14.7G**

Table V shows the input image sizes, and the numbers of parameters and the floating point of operations (FLOPs) of ViT [10], Swin transformer [41] and the proposed CSF transformer. Because of the design of the cross-fusion modules and shared-weights scheme, the proposed method has fewer number of parameters and FLOPs compared with ViT and Swin transformer. Based on the experimental results, the proposed CSF transformer not only outperforms the state-of-the-art transformers but also is more effective and efficient in memory usage and computation.

C. Ablation Study

To evaluate the proposed schemes, we adopt the BreakHis dataset which contains different magnification factors. The results of the ablation study are shown in Table VI. The first row shows the results of the proposed method without (w/o) using the multiple field-of-view (MFOV) patch embedding module, i.e. only a single field-of-view patch embedding is used to extract the transformer features. The results significantly degrade due to the lack of multiple field-of-view information and cross-scale information.

A naive idea is to apply conventional multiple-scale network to learn the representations of cells of difference scales without the proposed cross-scale fusion modules. The results without (w/o) the cross-fusion modules (CFMs) are shown in the second row of Table VI. In this case, the features of the last level of the transformer encoders are naively concatenated and passed to the channel-wise average pooling layer and a fully connected layer for prediction. Without the collaborations of features of different field-of-views, the transformer cannot well represent the cells for classification and performs even worse results compared with the results by using only patch embeddings of a single field-of-view. Such results reveal the importance of the proposed CFMs to fuse features from different field-of-views and the transformer encoder can then effectively learn the cross-scale patch embeddings by using the multi-head self-attention to represent cells and tissues of different field-of-views and scales.

To reduce the memory usage, avoid overfitting, and enforce the transformer encoders to learn features which are able to simultaneously represent cells of different field-of-views, we apply the shared-weights scheme. As shown in the third row of Table VI, the results without the shared-weights scheme also drop compared with the proposed method. In addition, the number of parameters of the network without the shared-weights scheme is about 93 M, while the number of parameters of the proposed method is only 57 M. Thus, the shared-weights scheme provides the benefits of the performance and memory usage.

In our method, we fuse the features of the larger field-of-views with the features of smaller field-of-views to assist the feature learning of the transformer encoders of the smaller field-of-views as shown in Fig. 3(a). To maintain the dimension consistency of features of the upsampling cross-fusion module, upsampling by using pixel shuffle is performed to features of the larger field-of-views. In this way, features of larger field-of-views provide more global context information to guide the learning of the features of the smaller field-of-views by using the multi-head self-attention schemes in the transformer encoders. However, two different cross-fusion modules shown in Fig. 3(b) and (c) can also be considered. Fig. 3(b) shows the structure of the downsampling cross-fusion module which assists the feature learning of the transformer encoders of the larger field-of-views. Fig. 3(c) shows the structure of the two-side cross-fusion module which assists the feature learning of the transformer encoders of the larger field-of-views and smaller field-of-views simultaneously. The results of the downsample cross-fusion and two-side cross-fusion modules are shown in the fourth and fifth rows of Table VI, respectively. These two cross-fusion modules achieve worse results because downsampling of smaller field-of-view features lead to information loss which degrades the learning of the transformation encoders.

Besides pixel shuffle, pyramidal concatenation [55] also integrates features of the low magnification factor with those of the high magnification factor by concatenation. The results of the

TABLE VI
ABLATION STUDY OF THE BREAKHIS DATASET BY USING RECOGNITION RATES (%)

Method	Image Level				Patient Level			
	40×	100×	200×	400×	40×	100×	200×	400×
w/o MFOV	93.47	96.03	96.45	93.99	94.10	95.11	96.03	93.01
w/o CFM	90.03	90.58	91.94	89.05	91.63	88.91	92.31	88.34
w/o Shared-Weights	97.42	98.68	99.19	97.53	98.00	98.57	99.20	97.72
Downsampling Cross-fusion	97.94	97.85	98.06	97.70	98.15	97.87	98.02	97.69
Two-side Cross-fusion	97.42	98.51	99.19	98.59	98.52	98.07	99.07	98.22
w Pyramidal Concatenation [55]	97.25	97.85	98.87	98.23	97.78	97.92	98.69	98.55
Proposed with $P = 32$	91.92	94.71	95.32	92.58	91.29	94.24	95.22	93.13
Proposed with $P = 16$	**98.11**	**99.01**	**99.52**	**99.47**	**98.89**	**99.22**	**99.58**	**99.46**

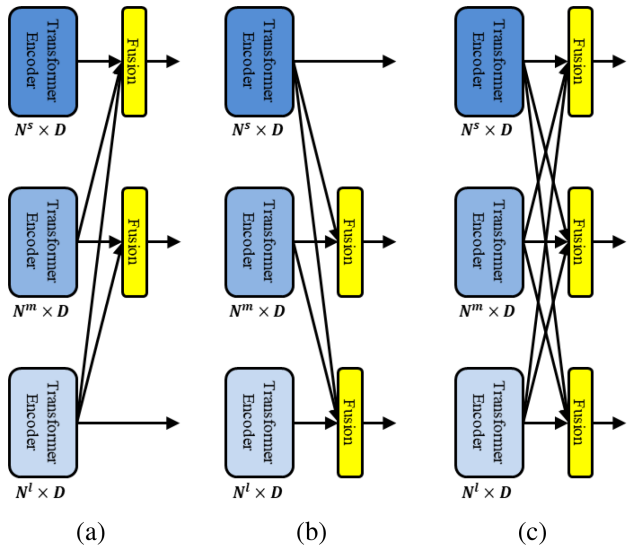

Fig. 3. Different cross-scale fusion modules. (a) The proposed upsampling cross-fusion module to assist the feature learning for the smaller field-of-views, (b) the downsampling cross-fusion module to assist the feature learning for the larger field-of-views, and (c) the two-sided cross-scale fusion module to assist the feature learning for the larger field-of-views and smaller field-of-views simultaneously.

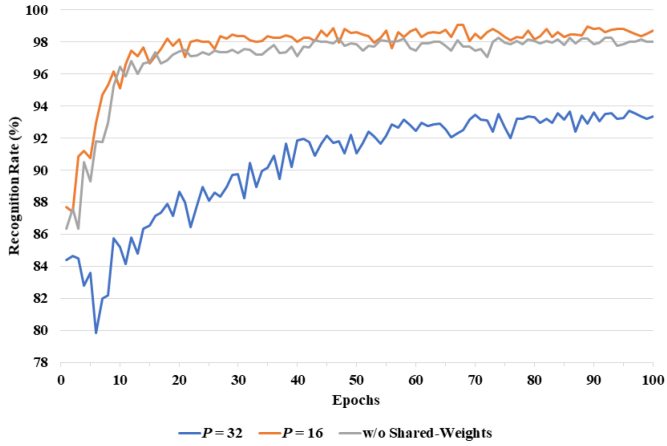

Fig. 4. Average image level recognition rates of the proposed method with the patch sizes 32, 16, and without (w/o) shared-weights with respect to the number of epochs in the BreakHis dataset.

proposed method with pyramidal concatenation are shown in the sixth row of Table VI. Because pixel shuffle uses channel-wise features with the sub-pixel convolution to convert low resolution feature maps to high resolution feature maps, the interpolated features are more robust to represent the patches of different field-of-views. Thus, the proposed method with pixel shuffle can achieve better results. Based on the ablation study, we show the effectiveness of the proposed method.

In the ablation study, we also compare the performance of the proposed method with different patch sizes. The seventh row in Table VI shows the results of the proposed method with the patch size 32. Because the patch size increases, the number N^i of the patches with respect to the field-of-view i will decrease. In this setting, the number of patches with respect to the large field-of-view is four. Transformer encoders can only discover rough contextual correlations among these four patches by using multi-head self-attention. Thus, the performance of the proposed method with the patch size 32 is worse than that of the proposed method with the patch size 16. The results show that setting a proper patch size with respect to the image resolution helps learn features to represent cells in the histopathological images.

The average image level recognition rates of the proposed method with the patch sizes 32 and 16 with respect to the number of epochs in the BreakHis dataset are shown in Fig. 4. The proposed method can achieve convergence in the target number of epochs for both patch sizes. Also shown in Fig. 4, the proposed method without using shared-weights achieves worse performance during the training process because it needs to update more parameters which may cause overfitting of the training data more easily. In contrast, the shared-weights scheme enforces the transformer encoders to simultaneously learn features from different field-of-views and helps the proposed method achieve better results.

D. Discussion

To solve the histopathological image classification problem under different cell sizes and magnification factors, the proposed CSF transformer adopts the cross-scale fusion module to integrate representations of different field-of-views. Compared with state-of-the-art methods, spatial cross-scale contextual information between different histopathological regions is learned via the transformer encoders based on the patch embeddings of different field-of-views. The experimental results show the effectiveness and generalization ability of the proposed CSF transformer with respect to four different histopathological datasets of different tissues and diseases. As shown in Table V, although the proposed CSF transformer is more computational efficient

TABLE VII
RESULTS OF THE BREAKHIS DATASET BY USING THE CONFUSION MATRIX IN (%)

Method		40×		100×		200×		400×	
	Class	Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
Proposed (Cases)	Malignant	99.47 (378/380)	0.53 (2/380)	99.06 (422/426)	0.94 (4/426)	99.54 (434/436)	0.46 (2/436)	99.74 (386/387)	0.26 (1/387)
Proposed (Cases)	Benign	4.46 (9/202)	95.54 (193/202)	1.12 (2/179)	98.88 (177/179)	0.54 (1/184)	99.46 (183/184)	1.12 (2/179)	98.88 (177/179)

TABLE VIII
RECALL, PRECISION AND F1-SCORE RESULTS OF THE BREAKHIS DATASET IN (%)

	40×	100×	200×	400×
Recall	99.47	99.06	99.54	99.74
Precision	97.67	99.53	99.77	99.48
F1-score	98.57	99.29	99.66	99.61

and has fewer parameters compared with conventional transformer methods, all patches are still considered in the multi-head self-attention computation. To further reduce the computation complexity, deformable attention [56] can be considered.

Table VII shows the confusion matrix of the proposed method in the BreakHis dataset with respect to the image level recognition rates. The proposed method can achieve equally high image level recognition rates with respect to different magnification factors. We also show recall, precision, and F1-score values of the BreakHis dataset of the proposed method with respect to image levels in Table VIII. The proposed method still achieves good recall, precision, and F1-score values with respect to different magnification factors. Please kindly note that the patient recognition rate is computed based on the average of correctly classified images. Thus, image level recall, precision and F1-score are more representative for evaluating the proposed method.

V. CONCLUSION

We propose the CSF transformer to solve the challenging histopathological image classification problems. Compared with natural images, histopathological images contain cells of variant sizes and are scanned under different magnification factors. To address these issues, we first propose the multiple field-of-view patch embedding module which provides variant views of histopathology images based on patching embeddings of different field-of-views. Then, the cross-fusion modules integrate features of different field-of-views to provide cross-scale contextual correlations so that the multi-head self-attention layers of the transformer encoders can simultaneously learn spatial correlations of patches of different field-of-views. Such schemes allow the CSF transformer to focus on learning contextual information between cells of variant sizes and magnification factors for histopathological image classification. In the experiments, we show that the CSF transformer outperforms the task-specific histopathological image classification methods and state-of-the-art transformer methods in four popular histopathological image datasets of different organs and magnification factors. Moreover, compared with state-of-the-art transformers, the CSF transformer requires less memory usage and few FLOPs. In the future, we will extend the proposed transformer structure to solve the pathological image segmentation problem.

ACKNOWLEDGMENT

The authors would like to thank National Center for High-performance Computing (NCHC) for providing computational and storage resources.

REFERENCES

[1] P.-C. Chung, W.-J. Yang, T.-H. Wu, C.-R. Huang, and Y.-Y. Hsu, "Emerging research directions of deep learning for pathology image analysis," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2022, pp. 100–104.
[2] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. K. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 738–751, Mar. 2016.
[3] H. Miyoshi et al., "Deep learning shows the capability of high-level computer-aided diagnosis in malignant lymphoma," *Lab. Investigation*, vol. 100, pp. 1300–1310, 2020.
[4] S. Belharbi, J. Rony, J. Dolz, I. B. Ayed, L. Mccaffrey, and E. Granger, "Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 702–714, Mar. 2022.
[5] J. Lin et al., "PDBL: Improving histopathological tissue classification with plug-and-play pyramidal deep-broad learning," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2252–2262, Sep. 2022.
[6] Y. Chen et al., "Dual polarization modality fusion network for assisting pathological diagnosis," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 304–316, Jan. 2023.
[7] N. Hashimoto et al., "Case-based similar image retrieval for weakly annotated large histopathological images of malignant lymphoma using deep metric learning," *Med. Image Anal.*, vol. 85, 2023, Art. no. 102752.
[8] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, "Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12589–12598.
[9] N. Hashimoto et al., "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3851–3860.
[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.
[11] Y. Zou, S. Chen, Q. Sun, B. Liu, and J. Zhang, "DCET-Net: Dual-stream convolution expanded transformer for breast cancer histopathological image classification," in *Proc. Int. Conf. Bioinf. Biomed.*, 2021, pp. 1235–1240.
[12] S.-K. Huang, C.-R. Yu, Y.-S. Liao, and C.-R. Huang, "Evaluations of deep learning methods for pathology image classification," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2022, pp. 95–99.
[13] U. Srinivas, H. S. Mousavi, V. Monga, A. Hattel, and B. Jayarao, "Simultaneous sparsity model for histopathological image representation and classification," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1163–1179, May 2014.
[14] Y. Hou, "Breast cancer pathological image classification based on deep learning," *J. X-ray Sci. Technol.*, vol. 28, no. 4, pp. 727–738, 2020.
[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[16] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recognit. Lett.*, vol. 29, no. 11, pp. 1684–1693, 2008.
[17] C.-R. Huang, C.-S. Chen, and P.-C. Chung, "Contrast context histogram–an efficient discriminating local descriptor for object recognition and image matching," *Pattern Recognit.*, vol. 41, no. 10, pp. 3071–3077, 2008.
[18] J. C. Caicedo, A. Cruz, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Proc. Artif. Intell. Med.*, 2009, pp. 126–135.
[19] M. M. Dundar et al., "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1977–1984, Jul. 2011.
[20] R. Sarkar and S. T. Acton, "SDL: Saliency-based dictionary learning framework for image similarity," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 749–763, Feb. 2018.
[21] X. Li, H. Tang, D. Zhang, T. Liu, L. Mao, and T. Chen, "Histopathological image classification through discriminative feature learning and mutual information-based multi-channel joint sparse representation," *J. Vis. Commun. Image Representation*, vol. 70, 2020, Art. no. 102799.
[22] X. Li, V. Monga, and U. K. A. Rao, "Analysis–synthesis learning with shared features: Algorithms for histology image classification," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 4, pp. 1061–1073, Apr. 2020.
[23] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2016, pp. 2560–2567.
[24] Y. Song, J. J. Zou, H. Chang, and W. Cai, "Adapting fisher vectors for histopathology image classification," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2017, pp. 600–603.
[25] J. Li et al., "Breast cancer histopathological image classification based on deep second-order pooling network," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.
[26] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.
[27] M. Toğaçar, K. B. Özkurt, B. Ergen, and Z. Cömert, "BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer," *Physica A, Stat. Mechanics Appl.*, vol. 545, 2020, Art. no. 123592.
[28] W. Mi et al., "Deep learning-based multi-class classification of breast digital pathology images," *Cancer Manage. Res.*, vol. 13, 2021, Art. no. 4605.
[29] Y. Zou, J. Zhang, S. Huang, and B. Liu, "Breast cancer histopathological image classification using attention high-order deep network," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 1, pp. 266–279, 2022.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[31] M. Liu et al., "A deep learning method for breast cancer classification in the pathology images," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 5025–5032, Oct. 2022.
[32] Z. Yang, L. Ran, S. Zhang, Y. Xia, and Y. Zhang, "EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images," *Neurocomputing*, vol. 366, pp. 46–53, 2019.
[33] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[34] B. Xu et al., "Attention by selection: A deep selective attention approach to breast cancer classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1930–1941, Jun. 2020.
[35] K. C. Burçak, Ö. K. Baykan, and H. Uğuz, "A new deep convolutional neural network model for classifying breast cancer histopathological images and the hyperparameter optimisation of the proposed model," *J. Supercomputing*, vol. 77, no. 1, pp. 973–989, 2021.
[36] Q. Qi et al., "Label-efficient breast cancer histopathological image classification," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 5, pp. 2108–2116, Sep. 2019.
[37] C. Sun, A. Xu, D. Liu, Z. Xiong, F. Zhao, and W. Ding, "Deep learning-based classification of liver cancer histopathology images using only global labels," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1643–1651, Jun. 2020.
[38] Z. Shao et al., "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 2136–2147, 2021.
[39] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16144–16155.
[40] Z. Wang, L. Yu, X. Ding, X. Liao, and L. Wang, "Lymph node metastasis prediction from whole slide images with transformer-guided multiinstance learning and knowledge transfer," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2777–2787, Oct. 2022.
[41] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10012–10022.
[42] T. E. Tavolara et al., "Automatic generation of the ground truth for tumor budding using H&E stained slides," in *Proc. Med. Imag. Digit. Comput. Pathol.*, 2022, pp. 40–46.
[43] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[45] 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.
[46] L. E. Boucheron, B. S. Manjunath, and N. R. Harvey, "Use of imperfectly segmented nuclei in the classification of histopathology images of breast cancer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 666–669.
[47] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2366–2378, Nov. 2015.
[48] K. Sirinukunwattana et al., "Gland segmentation in colon histology images: The GlaS challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, 2017.
[49] T. Guha and R. K. Ward, "Image similarity using sparse representation and compression distance," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 980–987, Jun. 2014.
[50] R. Sarkar and S. T. Acton, "Slide: Saliency guided image dictionary and image similarity evaluation," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 216–220.
[51] T.-H. Lin, J.-Y. Jhang, C.-R. Huang, Y.-C. Tsai, H.-C. Cheng, and B.-S. Sheu, "Deep ensemble feature network for gastric section classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 1, pp. 77–87, Jan. 2021.
[52] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1568–1576.
[53] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 642–651.
[54] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
[55] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14318–14328.
[56] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4784–4793.