This ACCV 2022 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2022
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

PathTR: Context-Aware Memory Transformer for Tumor Localization in Gigapixel Pathology Images

Wenkang Qin[1⋆], Rui Xu[1⋆], Shan Jiang[2], Tingting Jiang[1], and Lin Luo[1,3⋆⋆]

[1] Peking University, Beijing, China
[2] Beijing Institute of Collaborative Innovation, Beijing, China
[3] Southern University of Science and Technology, Shenzhen, China
{qinwk,xurui}@stu.pku.edu.cn,jiangs@bici.org,{ttjiang,luol}@pku.edu.cn

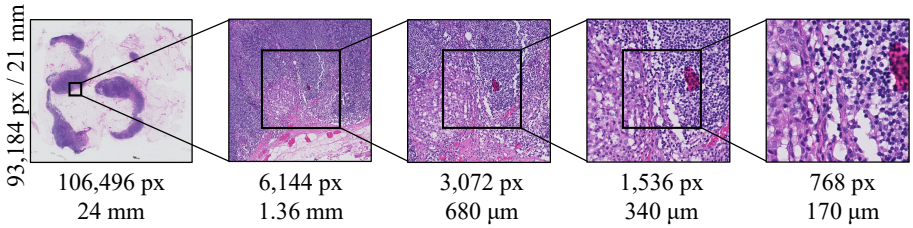
Abstract. With the development of deep learning and computational pathology, whole-slide images (WSIs) are widely used in clinical diagnosis. A WSI, which refers to the scanning of conventional glass slides into digital slide images, usually contains gigabytes of pixels. Most existing methods in computer vision process WSIs as many individual patches, where the model infers the patches one by one to synthesize the final results, neglecting the intrinsic WSI-wise global correlations among the patches . In this paper, we propose the PATHology TRansformer (PathTR), which utilizes the global information of WSI combined with the local ones. In PathTR, the local context is first aggregated by a self-attention mechanism, and then we design a recursive mechanism to encode the global context as additional states to build the end to end model. Experiments on detecting lymph-node tumor metastases for breast cancer show that the proposed PathTR achieves the Free-response Receiver Operating Characteristic Curves (FROC) score of 87.68%, which outperforms the baseline and NCRF method with +8.99% and +7.08%, respectively. Our method also achieves a significant 94.25% sensitivity at 8 false positives per image.

1 Introduction

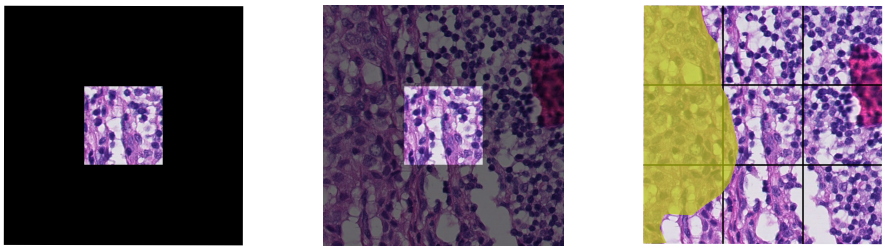
Pathology is the gold standard of clinical medicine, especially for cancer diagnosis. With the rapid development of digital slide scanners, digital pathology, where glass slides are digitized into whole slide images (WSIs), has emerged as a potential new trend. To distinguish the hierarchical morphological characteristics such as glands, cells, stroma, and nucleus, pathology slides are usually scanned with magnification at 200 multiple or 400 multiple, which produces extra large digital images of gigapixels.

Although the images can be stored and rendered in a multi-resolution pyramid manner, it brings considerable challenges to computer vision algorithms.

⋆ These authors contributed equally to this work.
⋆⋆ Corresponding author.



(a) A WSI is a very large gigapixel image with comprehensive pathological information that can be viewed in multiple scales. In a whole-slide-scale global view, it is difficult to recognize the details of tumor cells such as the shape of nucleus, which only presents a very small field of view at a high-resolution scale.



(b) Left: in a small patch that fits the regular neural network size (e.g. 224x224), although the details of nucleus are magnified can be viewed clearly, it is difficult to judge whether it contains tumor cells due to the lack of context patches surrounding it. Center: providing the context patches makes the task easier to detect tumor cells in the current patch. Right, the ground truth of the tumor area, which labeled as yellow.

Fig. 1: Overview of the challenges in tumor localization in gigapixel WSIs.

Taking a common task of tumor localization as an example, where the algorithm is to point out the suspicious locations in WSIs as boxes or heatmap, the input WSI usually is with a high resolution like 100, 000 x 100, 000 pixels. It is hard to infer such a high-resolution image by deep learning models.

There are several works[1,2,3,4,5] first divide WSIs into many patches. A deep learning model is used to classify patches, and these patch-level classification results are then organized into a heatmap to assist pathologists in tumor localization. The limitation of these methods is that the receptive field obtained by each patch is small, so the model may not be able to obtain enough spatial information (see Fig 1).

Some research works [6] referring to use both local and global spatial contexts only involve patch-level globel contexts, without utilizing *the spatial context information all-over the WSI* which reflects more structural disease characteristics. Some methods explore the effectiveness of local spatial information, such as [2] and its derivatives [3,4,5]. These methods aggregate some local patch information, and the results show that the model can obtain more accurate diagnosis results. In NCRF [2], neural conditional random fields are introduced to correlate the tumor probabilities of a central patch and its surrounding eight patches.

This method effectively improves the tumor detection results on WSI and obtains a smoother heatmap. Some other derivative works, such as [3,4], try to change the local patch of fixed position to the local patch of deformable position as deformable convolution did. In [5], Shen et al. explored the patch sampling strategy, and by modifying the patch sampling strategy, they obtained higher performance and faster inference speed on tumor localization. However, how the more global context can be exploited has not been explored.

Vision Transformers exhibit remarkable ability to reflect contextual relevance in computer vision area. By introducing a self-attention mechanism, different input tokens can perceive each other's information. Several works in Sec 2 use the Transformer to handle local and global contexts for video and language data. Inspired by them, our model utilizes Transformer to tackle with the large-scale WSI spatial context for tumor localization.

We proposed the Transformer-based model, PathTR, to combine local and global context within an end-to-end framework, especially to solve the large-scale context overflow issue. In our model, different patches' features are first extracted by the CNN backbone network. Then the features of a central patch and its surrounding features are further input into a Transformer encoder after adding positional encoding. At each layer of the Transformer Encoder, the context between different patches are accumulated through the self-attention layer. Through this simple approach, the local spatial context is more effectively utilized, and the tumor localization performance is effectively improved as described in Table 3. The next question is how to obtain larger spatial context information, even the spatial context information of the entire WSI. One of the simplest ways is that all the information on the entire WSI is input into the Transformer to obtain global perception. However, this method is difficult to implement because too many patches need to be input.

We further design a recursive mechanism to aggregate context over the entire WSI similar to RNN concept. During model initialization, as shown in Figure 2, we add several additional hidden states for global information aggregation in addition to the input patch features. These hidden states are designed as tokens of the same dimension as the input local contextual features. After each round of Transformer outputs, we update these hidden states to continuously aggregate the global context. Due to the introduction of the recursive mechanism, the order of patch input will affect the encoding of the hidden state. How serialize the patches on 2D space into a 1D sequence may affect the performance of the model. We further explored how the model's results are affected by different serialization methods, including row-wise, column-wise, and zigzag serialization. The results show that our model is robust to input order and achieves similar performance under different serialization methods.

We evaluated our model on the Camelyon 16 [7] tumor localization task, and the results show that our method significantly outperforms previous work, achieving FROC scores of 87.68%. It is worth noting that by introducing global context, our method can achieve 94.25% sensitivity under 8 average false posi-

tives per WSI. This result will be very beneficial for clinical applications, which are very sensitive to false positive numbers.

Overall, our main contributions are as follows:

1. We explored how to better utilize the local and global context information in WSI by introducing the self-attention mechanism and the recursive context management mechanism. The recursive mechanism encodes the local context into a hidden state, thereby obtaining the global WSI perception capability for the first time in the tumor localization task;
2. We explored the influence of input order, position encoding and other factors on this method, and the results show that our model is very robust to input order and other factors;
3. Our method achieves significant progress on the tumor localization task on the Camelyon 16 dataset, and reaches the state-of-the-art results. We hope our work can bring the clinical application of AI one step further in histopathology-assisted diagnosis.

2 Related Works

Tumor Localization Since IEEE International Symposium on Biomedical Imaging (ISBI) held the Camelyon challenge[7] in 2016, which first released a dataset of histopathological images with detailed annotations, there have been many excellent works trying to solve tumor localization and achieved good performance. Wang *et al.*[1] won the Camelyon 2016 championship, and then Liu *et al.*[8] from Google Brain achieved better performance under the same pipeline. And in [9], it was applied to the real world, and the possibility of its application in clinical practice was explored. Many subsequent methods used the same pipeline to explore under different settings and different tasks, for example. However, the pipeline used in the above method has the disadvantage of lack of context. Liu *et al.*[2] from Baidu Research used Conditional Random Field (NCRF) to explore spatial local context aggregation for the first time. Some follow-up work[10,11,12,13,14,15] used this method and explored this method on different tasks. Basically, these methods have not taken broader-concept context correlation into consideration, either in the context size or the feature-domain point of view.

Context Aggregation Context Aggregation has been widely used in many tasks in the field of large-volume text and video[2,10,11,13,16,17,18,19,20], and have achieved excellent performance. The tumor localization task in gigapixel pathology images explored brings new challenges of a large spatial size of contexts. In [21], Alexander *et al.*tried to use local and global to get global context of staining. Chomphuwiset *et al.*[22] uses Bayes networks to classify the patches around. In [23], the superpixel algorithm was used for segmentation and classification in low resolution as a global context. Our work analyze the characteristics of pathology images and introduce the framework to aggregate local and global morphological features as context, which has good potential to generalize.

3 Method

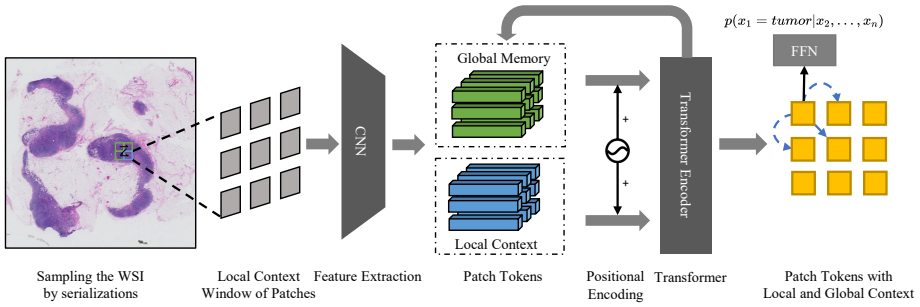


Fig. 2: **An overview of our proposed PathTR**. We sample WSIs into spatially adjacent patches, extract the feature representations of these patches through a convolutional neural network, and then perform context aggregation between patches through the Transformer[24]. We also save the features of all the inferred patches into the global memory. The global memory tokens also participate in the spatial context aggregation between patches to improve tumor localization performance.

In this section, we first describe the pipeline we use for tumor localization on WSIs, and focus on how PathTR improves the pipeline. In particular, we introduced how our proposed local context module and global memory module aggregate the local context and the global memory into different patches. The overall structure of the model can be represented as shown in Figure 2.

3.1 Preliminary

Pipeline of Tumor Localization Since most areas in the pathological images are background areas, and the background areas does not contain any tissue. We only randomly sample the normal patches and tumor patches from the foreground of the pathological image. To obtain the foreground mask, we use Otsu's method[25]. After obtaining the foreground mask, some points are sampled in the foreground area and use these points as the center point to crop out some patches in the normal area and the tumor area, and then train a binary classifier to diagnose the patches.

In the test phase, the trained model is tested by the sliding window manner on the foreground of the WSIs in the test set to obtain the tumor probability of each patch, and organize the probabilities of all patches into a heatmap as the output.

Problem Formulation The above classifier treats the tumor probabilities of different patches as independent of each other. So independently calculates the probability of each patch.

$$p(x_i = tumor) = f(x_i), x_i \in \mathbb{X} \tag{1}$$

where $p(x_i = tumor)$ represents the probability that a patch is a tumor, and $\mathbb{X}$ is the set of all patches. This inductive bias, which assumes that all patches are independent, is not that reasonable in pathological images, because whether each patch is a tumor is not only related to the current patch, but also to the surrounding patches, as shown in the Figure 1b. In [2], which introduces a neural conditional random field after CNN to correlate the context of P patches around a patch at probability level. They try to fit the conditional probability function Eq. (2).

$$p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_P) = f(x_1, \ldots, x_P), \\ P = p, x_1, \ldots, x_P \in \mathbb{X} \tag{2}$$

This enables some local context information, there are two issues in doing so. First, the window size of the local context P is difficult to determine in advance. If an excessive P is introduced, it will cause the model to be unable to infer due to hardware limitations. If the introduced P is too small, it will lead to inaccurate results because of a lack of context. Secondly, it performs context post-fusion after obtaining the probabilities of each patch *at the probability level*, which loses a lot of information about the patches' features.

To alleviate these two issues, we try to introduce local and global contexts at the feature level for feature aggregation. That is to say, our goal is to try to fit the Eq. (3).

$$p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = f(x_1, \ldots, x_n), \\ x_1, \ldots, x_n \in \mathbb{X} \tag{3}$$

3.2 Local and Global Context Aggregation

Different from the traditional pipeline sampling process, some non-adjacent patches may be randomly sampled. In PathTR, some windows are sampled in WSIs, and each window contains P patches (specifically, $P = 9$ in our experiment). When training, each sample is a window instead of a patch, that is to say, PathTR input $x \in \mathbb{R}^{N \times P \times C \times W \times H}$, where N, C, W, and H represent batch size, number of channels, width and height respectively. By this way, we can easily introduce local context. Similar to recurrent neural network (RNN), we retain the features of all the inferred patches into the global memory module to obtain a larger context. By introducing local context and global memory mechanism, we have solved the problems faced by models such as NCRF[2] above.

Local Context Aggregation Features extract network backbone is utilized, denoted as $f_{feat}(x) : \mathbb{R}^{N \times C \times W \times H} \to \mathbb{R}^{N \times M}$, to extract the features of these images, where M is the feature dimension. When patches are in f_{feat}, there is no correlation between different patches of different batches. Technically, x will be reshaped into $x \in \mathbb{R}^{NP \times C \times W \times H}$. After obtaining the features of all patches, we use the features aggregation network, denoted as $f_{aggr}(x) : \mathbb{R}^{N \times M} \to \mathbb{R}^{N \times M}$ to aggregate the features of different patches in the local context.

The features obtained by the P patches of the ith inference as an input tokens sequence, and add positional encoding,

$$z^i = [f_{feat}(x_1^i); f_{feat}(x_2^i); \ldots; f_{feat}(x_P^i)] + E_{pos}, \\ f_{feat}(x_p^i), E_{pos} \in \mathbb{R}^{P \times M}, \tag{4}$$

input it into the feature aggregation network $f_{aggr}(x)$, which is a Transformer encoder[24] in our implementation. For each Transformer encoder layer, it is aggregated of multi-head self-attention (MSA) and multilayer perceptron (MLP), and uses layer normalization to normalize the intermediate results, as described in Eqs. (5) to (7). The self-attention in each layer provides the ability to aggregate the local context as Eq. (5).

$$h_i = MSA(LN(z_{i-1})) + z_{i-1}, \qquad i = 1 \ldots L \tag{5}$$

$$l_i = MLP(LN(h_i)) + h_i, \qquad i = 1 \ldots L \tag{6}$$

$$f_{aggr}(z^i) = l_L, \tag{7}$$

where $z^i = [z_1^i; z_2^i; \ldots, z_n^i]$, and finally output the fused features $f_{aggr}(x) \in \mathbb{R}^{NP \times M}$. Refer to [24] for details of Transformer.

Global Context Aggregation The local context size P is required to determine whether a patch is a tumor is difficult to determine in advance. In order to make full use of the context, we have introduced a global memory module to record and aggregate the global contexts continuously. In pathological images, it is unrealistic to increase the size of the local context P unlimitedly, since its size is limited by the hardware. The local context in pathological images is far more important than the global context, but the global context has a role that cannot be ignored. Because global information may describe the overall information such as tissue and stain distribution of WSI. We save the global context by storing the inferred local context in the global memory module. That is, after the local context of the current patch is inferred by the model, the output tokens are encoded into the global memory module and wait to participate in the follow-up inferences. At the time of the ith inference, the information of $(i-1)P$ patches has been saved in the global memory module. With the patches in the local context module, a total of iP features of patches will be involved. This process can be formalized as follows:

$$y_i, z_{global}^{i+1} = f_{aggr}(z^i) = f_{aggr}([z_{local}^i; z_{global}^i]), \tag{8}$$

where $z_{local}^i = [z_1^i; z_2^i; \ldots, z_n^i]$, and $y_i, z_{global}^{i+1} \in \mathbb{R}^{NP \times M}$. Finally, we use a linear classifier to classify each patch embedding y_i.

The local context module and global memory module together constitute the core of PathTR. In order to illustrate how PathTR aggregate patches in different windows, context are progressively aggregated through the attention mechanism. For the patches currently input into the local context module, first use the feature network $f_{feat}(x)$ to get their feature tokens, and then aggregate through self-attention in Transformer[24]. Finally, the tokens in the local context module will aggregate the information in all past patches encoded in the global memory.

3.3 Robustness

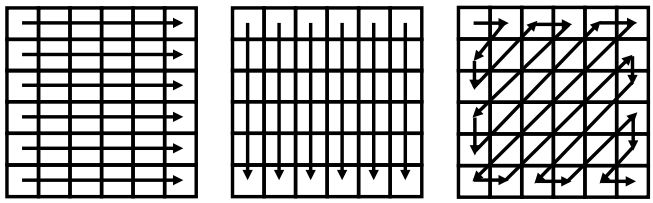
Since our model uses self-attention and recursion, our model may be sensitive to positional encoding or input order of images. We design different positional encodings and image serializations to verify the robustness of our model.

Positional Encoding Since Transformer[24] is not sensitive to spatial location, we have adopted the method of adding positional encoding to solve this problem as the convention. We tried three different position encoding methods to test our model. The first is the simplest version, which does not add any positional encoding. The second is to set a learnable parameter $E_{pos} \in \mathbb{R}^{P \times M}$. The third follows [24], we use sine positional encoding to encode, as follows:

$$E_{pos}(pos, 2i) = sin(\frac{pos}{temp^{2i/d_{model}}}) \tag{9}$$

$$E_{pos}(pos, 2i+1) = cos(\frac{pos}{temp^{2i/d_{model}}}) \tag{10}$$

Serialization The windows have to be converted into a sequence to input into PathTR. Obviously, the aggregation context in the global memory module is related to the input order of patches. In order to ensure that the most relevant context is aggregation in the global memory module as much as possible when judging whether a patch is a tumor, we tested three serializations to input windows into PathTR, including row-wise serialization, column-wise serialization, and zigzag serialization. Row-wise serialization and column-wise serialization respectively represent the input of sampled windows into PathTR row by row and column by column. In order to make every time input into the global context in PathTR center around the local context as much as possible, we imitated the JPEG encoding process[26] and adopted zigzag serialization to achieve this goal. This method ensures spatially adjacent patches are still as close as possible after the serialization. As shown in Figure 3.



(a) Row-wise Serialization. (b) Column-wise Serialization. (c) Zigzag Serialization.

Fig. 3: Serialization.

4 Experiments

4.1 Dataset

We conducted the experiments based on the Camelyon 16 dataset[7], which includes 160 normal and 110 tumor WSIs for training, 81 normal and 49 tumor WSIs for testing. † Table 1 describes the distribution of the Camelyon 16 WSIs. All WSIs were annotated carefully by the experienced pathologist, from which we can get pixel-level ROIs from the annotation mask. We conducted all the experiments on the largest scale, 40X magnification. Otsu algorithm[25] had been applied to exclude the background regions of each training WSI. Following the setting in [2]. We just randomly selected foreground patches during the training stage. Normal_001 to Normal_140 and Tumor_001 to Tumor_100 were selected for training, while other WSIs in the rest of the training set was used for validation. We also applied hard negative samples mining to select more patches from the tissue boundary regions as [2].

Table 1: Number of WSIs in the Camelyon 16 dataset[7]. Tumor means the number of slides including tumor regions in the training set. And Normal means the slides without any tumor region in the training set. Two slides in the test set are excluded because of the errors of annotations following [8]. So there are only 128 slides will be used in test set.

Institution	Tumor	Normal	Test
Radboud UMC	90	70	80
UMC Utrcht	70	40	50
Total	160	110	130

4.2 Implementation Details

We implement PathTR with PyTorch-1.8.0 and train the model with NVIDIA GeForce GTX 1080 Ti GPU. As our implementation is based on the open-source codebase[2], the methods such as patches generation and non-maximum suppression are similar to the NCRF[2]. At training time, we fetch 768×768 pixel windows from the training set, which are cropped as 3×3 grid of 256×256 pixel patches to feed the ResNet[27] backbone during the forward propagation. We train with Adam with a weight decay of 10^{-4} and initial learning rates of 10^{-3}. Our Transformer model is loaded with pre-trained weights from [28]. As shown in the Table 4, we report results with two different backbones: a ResNet-18 and a ResNet-34[27].

The Transformer encoder[24] is trained with a default dropout of 0.1. At training time, we try to select numbers of Transformer[24] encoder layers as 6 for default. And we compare the performance with sine, learned, and none

† The need for informed consent was waived by the institutional review board of Radboud University Medical Center (RUMC).[7]

positional encoding in ablation experiments. At the test stage, we use variant window grid size (baseline is 3×3) to aggregate the context at different scales of the surrounding regions. And we apply three types of serialization methods as described in Section 3.3, which are referred to as row-wise, column-wise and zigzag serialization.

In the ablation experiments, we use a training schedule of 20 epochs with a learning rate drop by a factor of 10 after 10 epochs, where a single epoch is a pass over all training patches once. Training the baseline model for 20 epochs on two 1080Ti GPUs takes about 24 hours, with 5 patches per GPU (hence a total batch size of 10). PathTR takes 0.01s per patch and about one hour per WSI at the inference time. So the total inference time of 130 WSIs in the test set is about 5 days using just one 1080Ti GPU.

4.3 Evaluation

Besides comparing the average accuracy and AUC with other methods, we also adopt the two important metrics, Free-response Receiver Operating Characteristic Curves (FROC) and sensitivity@nFP in the performance evaluation, because in clinical diagnosis the false negative rate is worth more attention [29].

The calculation of FROC score[29] is similar to that of Area Under Curve (AUC). We need to report the coordinates and confidence of tumors. If the coordinates are not in any tumor, it is judged as a false positive. If the reported coordinates successfully hit a tumor, it is considered as a successful judgment that a certain tumor exists. We can get the number of false positives and the tumor recall rate under different confidence thresholds. The FROC score[29] then be defined as the average sensitivity in the case of an average of 1/4, 1/2, 1, 2, 4, and 8 false positives for all WSIs on the test set.

4.4 Main Results

The proposed PathTR achieves the accuracy and AUC on par with other the state-of-the-art methods (Table 2). We further evaluate the FROC scores[29] of PathTR with local context module and global memory module setups and compare them with that of the baseline and of NCRF[2]. In addition, the test time augmentation is taken to improve the FROC score of our model on the test set following NCRF[2]. That is, in the test stage, the input patch is flipped or rotated, then PathTR is used for inference on augmented patches, and finally the multiple probability values are averaged to obtain the final tumor probability. Due to the introduction of spatial context, our method reduces false positive regions well, our method achieves a significant improvement in FROC score.

Table 3 shows the comparison with the Vanilla Pipeline and NCRF[2]. Our FROC score reaches 87.68% with test time augmentation and 94.25% sensitivity at 8 false positives per WSI. For comparison, a human pathologist attempting exhaustive search achieved 73.2% sensitivity.[8]

In order to compare the results from different models, the FROC curves of baseline, NCRF[2] and PathTR are presented in Figure 4a. In Figure 4b, the

Table 2: Performance Comparison of ACC and AUC.

Methods	ACC(%)	AUC
Baseline	96.79	0.9435
NCRF[2]	97.97	0.9725
Google[8]	-	0.9670
TransPath[6]	89.91	**0.9779**
PathTR	**98.19**	0.9757

improvement brought by different modules in PathTR are shown. With any false positives, the sensitivity has improved after the introduction of our local context module and global memory module.

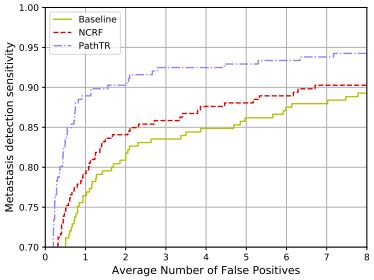
Table 3: **Performance comparison with the state of the art.** We test the results of adding a local context module and a global memory module to the baseline. At the same time we use test time augmentation to get better performance. We show the sensitivity achieved by our model at different false positives, as well as the FROC score. All models use grid size of 3×3, sine positional encoding, and row-wise input sequence, and with 6-layer Transformer encoder[24].

Methods	Local	Global	Sensitivity (%) @.25FPs	@.5FPs	@1FP	@2FPs	@4FPs	@8FPs	FROC (%)
Vanilla Pipeline[2]			66.98	71.90	77.20	81.80	84.95	89.28	78.69
NCRF[2]	✓		68.14	74.33	79.20	84.07	87.61	90.27	80.60
Wang *et al.*[1]			77.3	77.8	81.3	82.7	82.7	82.7	80.74
MSC-Net[4]	✓		-	-	-	-	-	-	80.78
DCRF[3]	✓		-	-	-	-	-	-	80.17
DP-FTD[5]	✓		-	-	-	-	-	-	81.7
DCRF-FTD[5]	✓		-	-	-	-	-	-	82.1
Ours (without global context)	✓		72.57	80.53	86.73	88.94	91.15	92.92	85.47
Ours	✓	✓	**76.55**	**83.63**	**88.94**	**90.27**	**92.48**	**94.25**	**87.68**

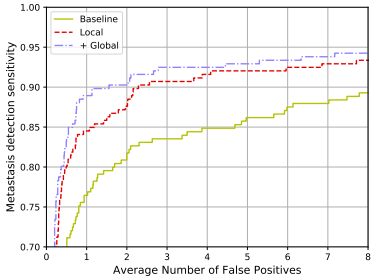
4.5 Ablation Study

Local and Global Module The local context module and global memory module are the core modules of PathTR aiming to aggregate a larger context, The local context module contains some spatially adjacent patches and completes the context aggregation in the Transformer[24]. The global memory module implicitly encodes the information of all past patches. we conduct ablation experiments in order to investigate the necessity of local context module and global memory module with different backbone networks.

As shown in Table 3, by introducing the local context module, the FROC is raised from the baseline 78.69% up to 84.59%, which is an improvement of 5.90%, proving the important role of local context feature-level fusion. Then introducing the global memory module receives a further improvement of 1.47% and reaches 86.06%. It demonstrates that global memory module can also contribute further improvement with a careful design.



(a) FROC of baseline, NCRF and PathTR results.



(b) FROC comparison among baseline, our local context module and our global memory module.

Fig. 4: FROC of baseline, with local context and global memory module.

We also explored whether the local context module and global memory module can achieve consistent improvements under different backbones. As shown in Table 4, we use ResNet-18 and ResNet-34 as the backbone.

Transformer Layers For aggregating the features from all patches of each window, we have taken Transformer encoder[24] to aggregate the local context and global memory. We set Transformer layers[24] from 2 to 8 with the interval of 2, since the experimental is too expensive for us. Table 5 shows the sensitivity at 8 false positives and FROC with different numbers of layers.

Table 4: Performance of global and local modules with different patch feature extraction backbone. Sensitivity is shown at 8 false positives per slide (same below). The FROC of baseline is reproduced by us, which is 78.25% in [2]

Methods	Backbone	Sensitivity (%)	FROC (%)
baseline	ResNet-18	89.28	78.69
+ Local	ResNet-18	93.36	84.59
+ Global	ResNet-18	93.81	86.06
baseline	ResNet-34	-	74.44
+ Local	ResNet-34	91.40	84.54
+ Global	ResNet-34	91.59	82.74

Running time and computational cost We test the inference times and FLOPs(on Nvidia GTX 1080Ti) of the Baseline, NCRF and PathTR to compare the computation overhead, as shown in Table 6. All input sizes are fixed as $9 \times 3 \times 224 \times 224$ (9 patches with 224×224 pixels).

Context Size The implementation of PathTR determines that the size of local context and global memory are identical. We call this as context size. In PathTR, the capacity of local context depends on the context size we take (in other words, the number of grids). The feature space of global memory will also

Table 5: Performance comparison with different numbers of Transformer encoder layers (using the same sine positional encoding and row-wise serialization).[24]

Methods	TF Layers	Sensitivity (%)	FROC (%)
Local	2	90.10	84.14
Local	4	90.97	84.07
Local	6	**93.36**	**84.59**
Local	8	89.39	82.70
+ Global	2	90.47	85.56
+ Global	4	91.03	85.74
+ Global	6	**93.81**	**86.06**
+ Global	8	93.36	84.36

Table 6: Speed Comparison.

Methods	Params (M)	Inference Time (patches/second)	GFLOPs
Baseline	11.18	144	16.367
NCRF	11.18	120	16.367
PathTR	30.09	73	16.709

increase with the increase of context size. With the increase of context size, there can be a wider context in the local context, and global memory can encode more global semantics. We tested three context sizes of 2×2, 3×3 and 3×6, as shown in Table 7. The results show that larger context size generates better but not significant performance gain.

Table 7: Performance of global memory module with different context size at test stage.

Methods	Context Size	Sensitivity (%)	FROC (%)
PathTR	2×2	91.79	85.20
PathTR	3×3	**93.81**	86.06
PathTR	3×6	92.18	**86.10**

Robustness Because of the recursive mechanism used in our model, all the local context is aggregated in global memory, and different input sequences may result in different results. Three serialization methods were used in our ablation experiments, as described in Section 3.3. The results are shown in Table 8. The three serialization methods achieved similar results, with Zigzag serialization slightly higher than row-wise and column-wise. This suggests that PathTR is not sensitive to input order, and that global memory plays a different role in the model than local context. Probably retaining more high-level semantic information, such as WSIs staining, instead of low-level semantic information, such as morphology.

Table 8: Performance of PathTR with variant serialization method. Zigzag serialization gains a slight FROC increment than other methods.

Methods	Serialization	Sensitivity (%)	FROC (%)
PathTR	Row-wise	**93.81**	86.06
PathTR	Column-wise	92.79	85.91
PathTR	Zigzag	93.80	**86.20**

The location of each patch in the local context is indispensable for clinical diagnosis, thus the order of tokens is needed to be fed into the local context and global memory tokens.

Two types of positional encoding are utilized in PathTR. Sine positional encoding is used to generate fixed position information. Learned positional encoding is added to allow the Transformer[24] to learn a set of appropriate positional information representations during the training process. The results are shown in Table 9.

Table 9: Performance of global and local modules with different positional encoding (all with 6-layer Transformer).

Methods	PE	Sensitivity (%)	FROC (%)
Local	None	90.76	82.20
Local	Learned	91.59	83.33
Local	Sine	**93.36**	**84.59**
Local + Global	None	91.50	83.30
Local + Global	Learned	92.04	**86.43**
Local + Global	Sine	**93.81**	86.06

5 Conclusion

This paper presents the PathTR method for tumor localization in gigapixel pathology images. We first introduce the Local Context Module to aggregate the local context surrounding a center patch. And then we bring in the global context of the whole slide images by introducing a recursive mechanism. The proposed PathTR can make full use of the locality of the image while retaining the global context, thus achieving a significant analysis capability of gigapixel images. We hope that our work can inspire more vision tasks that require analysis of gigapixel images to achieve better performance.

Acknowledgement This research was supported in part by the Foundation of Shenzhen Science and Technology Innovation Committee (JCYJ20180507181527806). We also thank Qiuchuan Liang (Beijing Haidian Kaiwen Academy, Beijing, China) for preprocessing data.

References

1. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. ArXiv preprint **abs/1606.05718** (2016) 2, 4, 11
2. Li, Y., Ping, W.: Cancer metastasis detection with neural conditional random field. ArXiv preprint **abs/1806.07064** (2018) 2, 4, 6, 9, 10, 11, 12
3. Shen, Y., Ke, J.: A deformable crf model for histopathology whole-slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2020) 500–508 2, 3, 11
4. Zhang, W., Zhu, C., Liu, J., Wang, Y., Jin, M.: Cancer metastasis detection through multiple spatial context network. In: Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition. (2019) 221–225 2, 3, 11
5. Shen, Y., Ke, J.: Sampling based tumor recognition in whole-slide histology image with deep learning approaches. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021) 2, 3, 11
6. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2021) 186–195 2, 11
7. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318** (2017) 2199–2210 3, 4, 9
8. Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al.: Detecting cancer metastases on gigapixel pathology images. ArXiv preprint **abs/1703.02442** (2017) 4, 9, 10, 11
9. Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D., Stumpe, M.C.: Artificial intelligence–based breast cancer nodal metastasis detection: insights into the black box for pathologists. Archives of pathology & laboratory medicine **143** (2019) 859–868 4
10. Shen, Y., Ke, J.: A deformable crf model for histopathology whole-slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2020) 500–508 4
11. Ye, J., Luo, Y., Zhu, C., Liu, F., Zhang, Y.: Breast cancer image classification on WSI with spatial correlations. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019, IEEE (2019) 1219–1223 4
12. Vang, Y.S., Chen, Z., Xie, X.: Deep learning framework for multi-class breast cancer histology image classification. In: International conference image analysis and recognition, Springer (2018) 914–922 4
13. Zanjani, F.G., Zinger, S., et al.: Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces. In: Medical imaging 2018: Digital pathology. Volume 10581., International Society for Optics and Photonics (2018) 105810I 4
14. Kong, B., Wang, X., Li, Z., Song, Q., Zhang, S.: Cancer metastasis detection via spatially structured deep network. In: International Conference on Information Processing in Medical Imaging, Springer (2017) 236–248 4

15. Mahbod, A., Ellinger, I., Ecker, R., Smedby, Ö., Wang, C.: Breast cancer histological image classification using fine-tuned deep network fusion. In: International Conference Image Analysis and Recognition, Springer (2018) 754–762 4
16. Oh, S.W., Lee, J., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE (2019) 9225–9234 4
17. Woo, S., Kim, D., Cho, D., Kweon, I.S.: Linknet: Relational embedding for scene graph. In Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., eds.: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. (2018) 558–568 4
18. Wu, C., Feichtenhofer, C., Fan, H., He, K., Krähenbühl, P., Girshick, R.B.: Long-term feature banks for detailed video understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE (2019) 284–293 4
19. Xu, J., Cao, Y., Zhang, Z., Hu, H.: Spatial-temporal relation networks for multi-object tracking. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE (2019) 3987–3997 4
20. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, IEEE (2020) 10334–10343 4
21. Wright, A.I., Magee, D., Quirke, P., Treanor, D.: Incorporating local and global context for better automated analysis of colorectal cancer on digital pathology slides. Procedia Computer Science **90** (2016) 125–131 20th Conference on Medical Image Understanding and Analysis (MIUA 2016). 4
22. Chomphuwiset, P., Magee, D.R., Boyle, R.D., Treanor, D.E.: Context-based classification of cell nuclei and tissue regions in liver histopathology. In: MIUA. (2011) 4
23. Zormpas-Petridis, K., Failmezger, H., Roxanis, I., Blackledge, M.D., Jamin, Y., Yuan, Y.: Capturing global spatial context for accurate cell classification in skin cancer histology. ArXiv preprint **abs/1808.02355** (2018) 4
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., eds.: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. (2017) 5998–6008 5, 7, 8, 9, 11, 12, 13, 14
25. Otsu, N.: A threshold selection method from gray level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9** (1979) 62–66 5, 9
26. Wallace, G.K.: The jpeg still picture compression standard. Commun. ACM **34** (1991) 30–44 8
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society (2016) 770–778 9
28. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby,

N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net (2021) 9
29. Egan, J.P., Greenberg, G.Z., Schulman, A.I.: Operating characteristics, signal detectability, and the method of free response. Journal of the Acoustical Society of America **33** (1961) 993–1007 10