GLOBAL-LOCAL ATTENTION NETWORK FOR WEAKLY SUPERVISED CERVICAL CYTOLOGY ROI ANALYSIS

Jun Shi[1], Kun Wu[1], Yushan Zheng[2, 4*], Yuxin He[1], Jun Li[3,4], Zhiguo Jiang[3,4], Lanlan Yu[5]*

[1] School of Software, Hefei University of Technology, Hefei, China
[2] School of Engineering Medicine, Beihang University, Beijing, China
[3] Image Processing Center, School of Astronautics, Beihang University, Beijing, China
[4] Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, China
[5] Motic (Xiamen) Medical Diagnostic Systems Co. Ltd., Xiamen, China

ABSTRACT

Existing supervised Convolutional Neural Network (CNN) approaches for cervical cytology image analysis generally rely on the heavy manual annotation for each cell or cell mass and thus lead to extensive time and effort. In this paper, we propose a global-local network for weakly supervised cervical cytology region of interest (ROI) analysis. It aims to perform the classification for ROIs and further classify the cells only with the ROI labels. Specifically, the proposed method firstly detects the cells within ROI and extracts the CNN features of cells. Then attention-based bidirectional LSTM (Att-BLSTM) is applied to explore the global contextual information within ROI. On the other hand, the Vision Transformer (ViT) is used to exploit the local attentive representations of the cells in ROIs. The cross attention (CA) is applied to incorporate the global contextual features and local patterns and thus generates more discriminative feature representation of ROI. More importantly, the CA score is used as the pseudo label to select top and least attentive cells. Therefore, the in-the-class and out-of-the-class CA branches are trained to achieve the cell classification. Experimental results demonstrate the effectiveness of our method for cervical cytology ROI and cell classification, and the weak supervision of the image-level label has great potential to promote the automatic whole slide cervical image analysis and alleviate the workload of cytologists.

Index Terms—Cervical cancer screening, cervical cytology ROI analysis, weakly supervised learning.

1. INTRODUCTION

As the most common screening of cervical cancer at early stage, cervical cytology screening requires cytologists to observe the morphological change of cells in the Papanicolaou (PAP) smears under the microscope, then identify the normal and abnormal cell types and make the final diagnosis according to The Bethesda System (TBS). However, the screening process is usually time-consuming and labor-intensive. Therefore, the automatic screening methods have been proposed to intelligently identify the abnormal cells in the cytology slide and thus reduce the workload of cytologists.

With the development of artificial intelligence (AI), CNNs achieve the promising performance in computer vision and has been gradually applied in cervical cytology screening. Zhang et al. [1] apply CNN to extracts deep features of cell image patch for cervical cell classification. Plissiti et al. [2] introduce the public cervical cell image dataset SIPaKMeD and use VGG to achieve the better classification performance compared with the handcrafted features. However, these CNN-based classification methods rely on heavy manual cell-level annotation. Therefore, it is likely to cost extensive time and effort and easily generate the noisy samples due to the subjective differences of cytologists. To address these problems, some studies [3, 4] model the screening process as the ROI classification instead of using fine-grained cell classification. It effectively overcomes the annotation difficulties and makes the AI-assisted screening more practically deployable. Concretely, Ferreira et al. [3] use the saliency map from cytologists' eye-tracking to detect candidate cells and then classify the ROIs. Gupta et al. [4] partition the WSI into fixed-size ROIs and perform three deep learning classifiers on ROIs for anomaly identification. Nevertheless, these above methods directly take the entire ROI as the training input and fail to explore the global contextual information within ROI. Recent studies [5, 6] show the methods based on Recurrent Neural network (RNN) can exploit the contextual relationship through modelling the ROI or WSI as a sequence of features. More importantly, the intrinsic relationship between ROI and local cells is not considered, which contributes to identifying the potential abnormal cells instead of using the elaborate manual annotation.

* Corresponding author: juns@hfut.edu.cn; yszheng@buaa.edu.cn

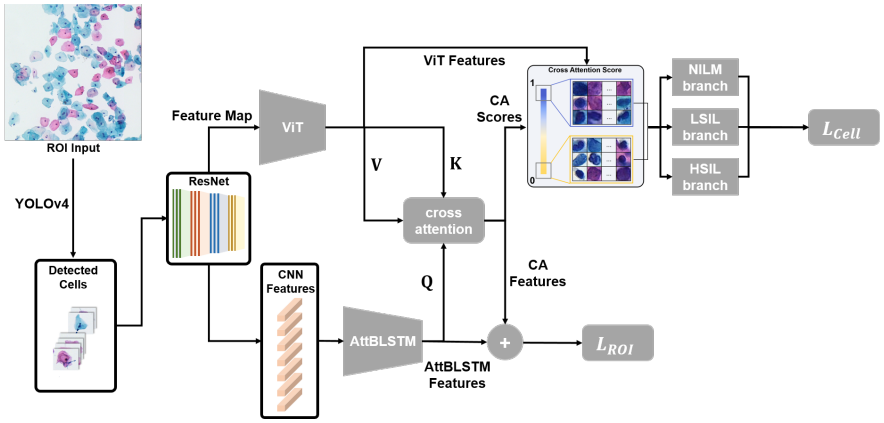

Fig. 1. Pipeline of our method for cervical cell image analysis.

In this paper, we propose a novel global-local attention network for weakly supervised cervical cytology ROI analysis. Unlike the above-mentioned methods, we achieve the ROIs and cell classification in a weakly supervised way. Only the ROI label is used in the feature learning and thus alleviate the workload of manual annotation for cells and avoid the subjective annotation differences of cytologists. Concretely the cells are firstly detected by YOLOv4 [7] pretrained in the cell detection task which only provides the location of cells and thus the corresponding CNN features can be gained. Then cell features are used as the sequential data and fed into the attention-based bi-directional LSTM (Att-BLSTM) [8] for capturing the global contextual information of ROI. Meanwhile, the Vision Transformer (ViT) [9] is performed on the feature maps of cells for exploring the local attentive representations. The cross attention (CA) is used to incorporate the global contextual features and local attentive features. Therefore, the more discriminative feature representation of ROI can be obtained for ROI classification. More importantly, CA score is taken as the pseudo label to select top and least attentive cells. Benefit from the weakly supervised setting of clustering-constrained-attention multiple-instance learning (CLAM) [10], we train the in-the-class and out-of-the-class CA branches with the selected cells for the cell classification. Furthermore, we present a liquid based cervical cytology dataset which includes 900 ROIs with 3 ROI-level labels (NILM, LSIL and HSIL). Experimental results on this dataset demonstrate the feasibility and effectiveness of our method for cervical cytology ROI and cell classification.

2. METHODOLOGY

2.1. Overview

The pipeline of the proposed method is illustrated in Fig. 1. The cervical cells are firstly detected by YOLOv4 and then the corresponding CNN features of cells extracted by ResNet-50 [11] are fed into Att-BLSTM for capturing the global contextual information within ROI. On the other hand, the feature maps of Stage 4 from ResNet-50 are inputted to the hybrid architecture of ViT pre-trained on the ImageNet-21k and thus the local attentive features of cells can be gained. Afterwards, the cross attention (CA) is applied to highlight the representative cells which builds the connections between the ViT features of cells and the AttBLSTM features of ROI. Specifically, the ViT features after linear transformation are respectively the key **K** and value **V** inputs of CA, and the AttBLSTM features after linear transformation are taken as the query **Q**. Consequently, the final ROI feature representation is generated by the sum operation of CA features and AttBLSTM features. Meanwhile, the N top and N least attentive cells are selected by the sorted CA scores, and respectively regarded as the positive and negative samples. Similar to CLAM [10], given the number of classes C, we construct C branches and each branch has a binary classification layer which predicts whether the sample belongs to this class and the corresponding probability. Note that the in-the-class CA branch corresponds the true label of given ROI and the remaining are the out-of-the class branches. For the in-the-class branch, the ViT features of N positive and N negative samples are all fed. Particularly, the ViT features of N positive samples are taken as the negative samples of the out-of-the class branch. The entire network is trained by a hybrid loss consisting of the cross-entropy loss L_{ROI} for ROI classification and the cell classification loss L_{cell} which has C class-specific cross-entropy loss functions.

2.2. Global attention branch

Inspired by the success of RNN-based sequence modelling on the ROI and WSI [5, 6], we firstly use pre-trained ResNet-50 to extract the features of detected cells and then model the cell features as a sequential data. Assuming n_c detected cells are represented by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_c}] \in \mathbb{R}^{d_f \times n_c}$ where d_f is the feature

dimension of ResNet-50. Att-BLSTM is performed on the sequential cell features $\mathbf{X}$ to obtain the ROI representation. The structure of LSTM is formulated as [8]:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i\right) \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{g}_t &= tanh\left(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{W}_{cc}\mathbf{c}_{t-1} + \mathbf{b}_c\right) \\
\mathbf{c}_t &= \boldsymbol{i}_t\boldsymbol{g}_t + \boldsymbol{f}_t\boldsymbol{c}_{t-1} \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o\right) \\
\mathbf{h}_t &= \mathbf{o}_t tanh\left(\mathbf{c}_t\right)
\end{aligned} \tag{1}$$

where $\mathbf{i}_t$ is the input gate with the corresponding weight matrix $\mathbf{W}_{xi}$, $\mathbf{W}_{hi}$, $\mathbf{W}_{ci}$ and $\mathbf{b}_i$, $\mathbf{f}_t$ is the forget gate with the corresponding weight matrix $\mathbf{W}_{xf}$, $\mathbf{W}_{hf}$, $\mathbf{W}_{cf}$ and $\mathbf{b}_f$, $\mathbf{o}_t$ is the output gate with the corresponding weight matrix $\mathbf{W}_{xo}$, $\mathbf{W}_{ho}$, $\mathbf{W}_{co}$ and $\mathbf{b}_o$, $\mathbf{c}_t$ is the current cell state, $\mathbf{h}_t$ is the hidden state of the current time step $t\left(\mathbf{h}_0 = \mathbf{0}\right)$ and σ denotes the sigmoid activation function.

Considering the bidirectional LSTM (BLSTM) has two sub-networks for two directions, the output of BLSTM is represented by $\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}_t} \oplus \overleftarrow{\mathbf{h}_t}\right]$ where $\overrightarrow{\mathbf{h}_t}$ and $\overleftarrow{\mathbf{h}_t}$ denote the output of the forward LSTM and backward LSTM respectively. To focus on the most important semantic information in the sequential data, the attention mechanism is used in Att-BLSTM and the final feature representation of ROI $\mathbf{h}_{ROI}$ can be gained by Eq. (2):

$$\boldsymbol{\alpha} = softmax\left(\mathbf{w}^{\mathrm{T}} tanh\left(\mathbf{H}\right)\right),\ \mathbf{h}_{ROI} = tanh\left(\mathbf{H}\boldsymbol{\alpha}^{\mathrm{T}}\right) \tag{2}$$

where $\mathbf{H}$ is the output matrix of BLSTM and $\mathbf{w}$ is the trained parameter vector. Consequently, the global contextual information within ROI can be captured.

2.3. Local attention branch

To exploit the local attentive representations of the cells in ROIs for improving the ROI feature representation ability and further achieve the cell classification in a weakly supervised way, ViT is performed on the detected cells. Instead of using the raw image patches, the feature maps of Stage 4 from ResNet-50 are fed into the hybrid architecture of ViT pre-trained on the ImageNet-21k. Namely, the patches can be regarded as have the spatial size 1×1 [9].

The ViT has L Transformer encoders which include multiheaded self-attention (MSA), Layernorm (LN) and MLP blocks, which can be characterized as follows [9]:

$$\begin{aligned}
&\mathbf{z}_0 = \left[\mathbf{b}_{class}; \mathbf{b}_1\mathbf{E}; \mathbf{b}_2\mathbf{E}; \ldots; \mathbf{b}_m\mathbf{E}\right] + \mathbf{E}_{pos}, \\
&\mathbf{E} \in \mathbb{R}^{d_c \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(m+1)\times D}
\end{aligned} \tag{3}$$

$$\mathbf{z}'_l = \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}_{l-1}\right)\right) + \mathbf{z}_{l-1}, l = 1, \cdots, L \tag{4}$$

$$\mathbf{z}_l = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{z}'_l\right)\right) + \mathbf{z}'_l, l = 1, \cdots, L \tag{5}$$

$$\mathbf{y} = \mathrm{LN}\left(\mathbf{z}_L^0\right) \tag{6}$$

where $\mathbf{b}_i\left(i = 1, \ldots, m\right)$ denotes the 1×1 patches, $\mathbf{z}_0^0 = \mathbf{b}_{class}$ is a learnable embedding to the input sequence, d_c is the channels of the feature maps, and $\mathbf{E}$ is the learnable embedding matrix which project the patches into the D dimensional latent representation of the Transformer, $\mathbf{E}_{pos}$ is the position embeddings, MLP has two layers with GELU activation function and $\mathbf{y}$ is the final representation of cell.

2.4. Global-Local feature learning

After the ROI feature $\mathbf{h}_{ROI} \in \mathbb{R}^{1024}$ extracted by the Att-BLSTM and the ViT features of cells $\mathbf{Y} = \left[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_c}\right] \in \mathbb{R}^{n_c \times D}\left(D = 768\right)$ are obtained, the cross attention (CA) is used to build the connections between the ViT features of cells and the AttBLSTM features of ROI, which can highlight the representative cells and thus improve the discriminant ability of ROI feature. Note that n_c denotes the number of detected cells. The cross attention can be formulated:

$$\begin{aligned}
\mathbf{a} &= Attention\left(\mathbf{h}_{ROI}\mathbf{W}^Q, \mathbf{Y}\mathbf{W}^K, \mathbf{Y}\mathbf{W}^V\right) \\
&= softmax\left(\frac{\left(\mathbf{h}_{ROI}\mathbf{W}^Q\right)\left(\mathbf{Y}\mathbf{W}^K\right)^{\mathrm{T}}}{\sqrt{d_k}}\right)\left(\mathbf{Y}\mathbf{W}^V\right)
\end{aligned} \tag{7}$$

where $\mathbf{W}^Q \in \mathbb{R}^{d_f \times d_k}$, $\mathbf{W}^K \in \mathbb{R}^{D \times d_k}$, $\mathbf{W}^V \in \mathbb{R}^{D \times d_v}$, 且 $d_f = d_k = d_v = 1024$. $\mathbf{a}$ is regarded as the weighted sum of the cell ViT features. The final ROI feature $\tilde{\mathbf{h}}$ can be gained by encoding the local cell features into the global contextual features, as shown in Eq. (8), and it can be used to train the cross-entropy loss L_{ROI} for classifying the ROIs.

$$\tilde{\mathbf{h}} = \mathbf{h}_{ROI} + \mathbf{a} \tag{8}$$

Furthermore, the CA score is defined as follows:

$$\mathbf{s} = softmax\left(\frac{\left(\mathbf{h}_{ROI}\mathbf{W}^Q\right)\left(\mathbf{Y}\mathbf{W}^K\right)^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{9}$$

which can characterize the attention responses of cells corresponding to the global ROI. Therefore, motivated by the weakly supervised setting of CLAM [10], the CA score is taken as the pseudo label to weakly supervise the cell classification. Concretely, N top and N least attentive cells are selected to be the positive and negative samples according to the sorted CA score. Then we build C branches and each branch has a binary classification layer which predicts whether the sample belongs to this class and the corresponding probability. Moreover, the class-specific cross-entropy loss function corresponding to each branch is used to train the binary classification layer. Note that we only consider 3 categories ($C = 3$) of cervical cell in this

Table 1. Comparison of ROI classification accuracies (%).

Methods	Gupta et al. [4]	CLAM [10]	RNN	GRU	LSTM	Att-BLSTM	Att-BLSTM +ViT w/o L_{cell}	Ours
Accuracy	62.70	65.56	61.48	65.92	68.15	71.85	72.59	**73.33**

paper, namely NILM, LSIL and HSIL. Similar to CLAM, we also define the in-the-class CA branch which corresponds to the true label of given ROI and the remaining are the out-of-the class branches. The in-the-class and out-of-the-class CA branches are trained to achieve the cell classification through the selected positive and negative samples. Considering different classes are mutually exclusive, the detailed sample selection rule is given:

1. For the NILM category, all the cell samples as the positive samples are fed into the NILM branch.

2. For the LSIL category, N positive and N negative samples are fed into the LSIL branch where $N = n_c \times 10\%$. Besides, the N positive samples are taken as the negative samples to input the NILM branch.

3. For the HSIL category, N positive and N negative samples are fed into the HSIL branch where $N = n_c \times 10\%$. The N positive samples are used as the negative samples to input the NILM and LSIL branches, respectively.

All the branches are trained through the selected samples and corresponding class-specific cross-entropy loss function. Therefore, the cell classification loss L_{cell} is formulated as $L_{cell} = \left(L_{NILM} + L_{LSIL} + L_{HSIL}\right)/3$ where L_{NILM}, L_{LSIL} and L_{HSIL} are the class-specific cross-entropy loss function. The network is trained by the loss $L = L_{ROI} + \lambda L_{cell}$ and λ is the regularization parameter.

3. EXPERIMENTS

To evaluate the effectiveness of our method, the liquid based cervical cytology dataset supplied by Motic is proposed which consists of 900 ROIs with 3 ROI-level labels (NILM, LSIL and HSIL) shown in Fig. 2. Each category has 300 ROIs. The size of ROI ranges from 1024×1024 to 2048×2048. Note that the label of ROIs is only available. The cell images detected by YOLOv4 are resized to 224×224 and represented by the $d_f = 2048$ dimensional ResNet-50 features. The network is trained for 90 epochs by stochastic gradient descent (SGD) and λ is set to 0.25. The initial learning rate is set to 0.0003 and divided by 10 for every 30 epochs. All the experiments are conducted on a computer with an Intel Core i7-7820X CPU of 3.60 GHz and 2 GPUs of NVIDIA GTX 2080Ti.

The ROI classification results of different methods are presented in Table 1. Note that the four methods (RNN, GRU, LSTM and Att-BLSTM) only use the global contextual features for ROI classification and BLSTM+ViT w/o L_{cell} encodes the ViT features into the Att-BLSTM but without the cell classification loss L_{cell}. As shown in Table 1, Att-BLSTM is better than the conventional ROI classification methods (Gupta et al. [4], CLAM [10], RNN, GRU and LSTM) since it uses the attention-based LSTM to capture the most informative contextual representation for the sequential features. As the ViT features of cells are encoded into the global Att-BLSTM features, the method BLSTM+ViT w/o L_{cell} is superior to Att-BLSTM. It indicates the local ViT features contributes to improving the representational ability of global ROI features. More importantly, the proposed method outperforms other methods. It can be explained that our method achieves the trade-off between ROI and cell classification in a weakly supervised way and thus yields more discriminant power.

Moreover, the cells within 900 ROIs are further annotated by the cytologists for the cell classification performance comparison of supervised CNN methods (DeepPap [1] and ResNet-50). Note that 70% of cells from each category are used to train and the remaining cells are taken as the test set. As shown in Table 2, our method is comparable with the supervised DeepPap and ResNet-50. It indicates the weakly supervised learning designed by our method is beneficial to cell classification and generates promising classification performance without much manual cell-level annotation.

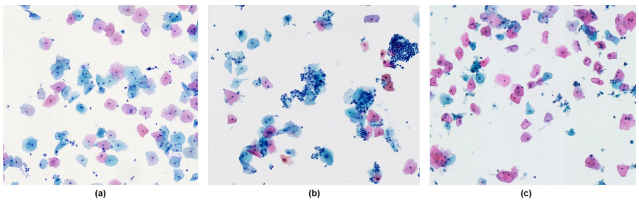

Fig. 2. ROIs of cervical cytology for 3 ROI-level categories: (a) NILM, (b) LSIL, (c) HSIL.

Table 2. Comparison of cell classification accuracies (%).

Methods	DeepPap [1]	ResNet-50 [11]	Our method
Accuracy	89.33	95.48	**93.53**

4. CONCLUSION

In this paper, we introduce a weakly supervised cervical cytology ROI analysis method based on global-local attention network. It respectively applies Att-BLSTM and ViT to explore the global contextual information within ROI and potential relationship between ROI and local cells. The cross attention (CA) is used to encode cell ViT features into the ROI representation. Consequently, the discriminative ability of ROI features is enhanced. More importantly, the CA score is taken as the pseudo label to weakly supervise the cell classification through the in-the-class and out-of-the-class branches. Experiments on the proposed cervical cytology ROI dataset demonstrate our method yield better ROI classification result and particularly shows the promising cell classification performance.

5. ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China (grant no. 61906058, 61901018, 62171007 and 61771031), partly supported by the Anhui Provincial Natural Science Foundation (grant no. 1908085MF210), partly supported by the China Postdoctoral Science Foundation (grant no. 2019M650446), and partly supported by the Fundamental Research Funds for the Central Universities of China (grant no. JZ2020YYPY0093).

6. REFERENCES

[1] L. Zhang, L. Lu, I. Nogues, et al., "DeepPap: Deep convolutional networks for cervical cell classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1633-1643, 2017.

[2] M.E. Plissiti, P. Dimitrakopoulos, G. Sfikas, et al., "SIPaKMeD: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images," in *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 3144-3148.

[3] D.S. Ferreira, G.L.B. Ramalho, F.N.S. Medeiros, et al., "Saliency-driven System with Deep Learning for Cell Image Classification," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, Venice, Italy, 2019, pp. 1284-1287.

[4] M. Gupta, C. Das, A. Roy, et al., "Region of Interest Identification for Cervical Cancer Images," in *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, Iowa City, USA, 2020, pp. 1293-1296.

[5] Y. Zheng, Z. Jiang, F. Xie, et al., "Diagnostic Regions Attention Network (DRA-Net) for Histopathology WSI Recommendation and Retrieval," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 1090-1103, 2021.

[6] S. Tripathi, S. K. Singh, and H. K. Lee, "An End-to-end Breast Tumour Classification Model Using Context-Based Patch Modelling – A BiLSTM Approach for Image Classification," *Computerized Medical Imaging and Graphics*, vol. 87, 101838, 2021.

[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv: 2004.10934*, 2020.

[8] P. Zhou, W. Shi, J. Tian, et al., "Attention-based Bidirectional Long Short-Term Memory Networks for Relation Classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany, 2016, pp. 207-212.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929*, 2021.

[10] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, et al., "Data-efficient and Weakly Supervised Computational Pathology on Whole-Slide Images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555-570, 2021.

[11] K. He, X. Zhang, S. Ren, et al., "Deep Residual Learning for Image Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 770-778.