SETMIL: Spatial Encoding Transformer-Based Multiple Instance Learning for Pathological Image Analysis

Yu Zhao[2], Zhenyu Lin[1,2], Kai Sun[2,3], Yidan Zhang[4], Junzhou Huang[1,2,3,4], Liansheng Wang[1(✉)], and Jianhua Yao[2(✉)]

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen 361005, China
lswang@xmu.edu.cn

[2] AI Lab, Tencent, Shenzhen 518000, China
jianhuayao@tencent.com

[3] School of Basic Medical Science, Central South University, Changsha 410013, China

[4] School of Computer Science, Sichuan University, Chengdu 610065, China

Abstract. Considering the huge size of the gigapixel whole slide image (WSI), multiple instance learning (MIL) is normally employed to address pathological image analysis tasks, where learning an informative and effective representation of each WSI plays a central role but remains challenging due to the weakly supervised nature of MIL. To this end, we present a novel Spatial Encoding Transformer-based MIL method, SETMIL, which has the following advantages. (1) It is a typical embedded-space MIL method and therefore has the advantage of generating the bag embedding by comprehensively encoding all instances with a fully trainable transformer-based aggregating module. (2) SETMIL leverages spatial-encoding-transformer layers to update the representation of an instance by aggregating both neighbouring instances and globally-correlated instances simultaneously. (3) The joint absolute-relative position encoding design in the aggregating module further improves the context-information-encoding ability of SETMIL. (4) SETMIL designs a transformer-based pyramid multi-scale fusion module to comprehensively encode the information with different granularity using multi-scale receptive fields and make the obtained representation enriched with multi-scale context information. Extensive experiments demonstrated the superior performance of SETMIL in challenging pathological image analysis tasks such as gene mutation and lymph node metastasis prediction.

Keywords: Multiple instance learning · Pathological image analysis · Transformer · Position encoding

Y. Zhao, Z. Lin, and K. Sun—Equally-contributed authors.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13432, pp. 66–76, 2022.
https://doi.org/10.1007/978-3-031-16434-7_7

1 Introduction

In modern healthcare, pathological image analysis plays a crucial role in the process of disease detection, interpretation, and is regarded as the gold standard for the diagnosis of almost all types of cancer [11,15,16,20,26,29]. The huge size of the WSIs draws a challenge for deep learning-based methods to comprehensively encode information of the entire WSI with conventional architecture. To tackle this issue, multiple instance learning (MIL) is usually leveraged to formulate pathological image analysis tasks into weakly supervised learning problems [1]. Generally, there exist two main categories of MIL methods in pathological image analysis, i.e., instance-space MIL and embedded-space MIL [1,23]. Instance-space MIL methods usually focus their learning process on the instance level and often achieve inferior performance compared to other MIL methods [1,8,9]. By contrast, embedded-space MIL methods attempt to extract information globally at the bag level, which can comprehensively exploit the entire WSI in pathological image analysis.

In embedded-space MIL methods, instances are firstly embedded into low-dimensional representations and then integrated to format a bag-level representation by an aggregating module, therefore transforming the multiple instance learning problem into a standard supervised learning problem [3,13,18,19]. Thus, developing an effective aggregating module to generate bag-level representation is a key step in embedded-space MIL methods, which remains a challenging problem for the following reasons. First of all, conventional fixed or parameterized pooling-based bag-embedding methods [23,25,30] are either fixed or partially trainable and therefore has limited ability to represent the bag information. Second, state-of-the-art attention-based MIL methods [6,7,13] represent the bag as a weighted sum of instance features, which is just a linear combination rather than a high-level feature embedding. Besides, these attention-based MIL methods lack sufficient consideration of position and context information of tiled patches (instances) in the WSI (bag). Third, recurrent neural network (RNN)-based MIL [2] or pioneering work employing transformer in image analysis [4] has considered position and context information. However, how to effectively and efficiently encode 2D positions of patches inside a WSI in the one-dimensional-sequence architecture is still an open question.

In this work, we aim at developing a MIL method to solve challenging pathological image analysis tasks, especially those needing comprehensive consideration of the tumour micro-environment on the entire WSI, such as metastasis and gene mutation prediction [5]. Taking into account that pathologists usually make diagnoses by leveraging both the context information locally around a single area and the correlation information globally between different areas, the developed MIL model should try to mimic this clinical practice. Therefore, we propose to update the representation of an instance by aggregating representations of both neighbouring instances (local information, a sub-region of WSI should have similar semantic information) and globally corrected instances simultaneously (similar sub-regions have similar semantic information). To summarize, we present a novel Spatial Encoding Transformer-based Multiple Instance Learning

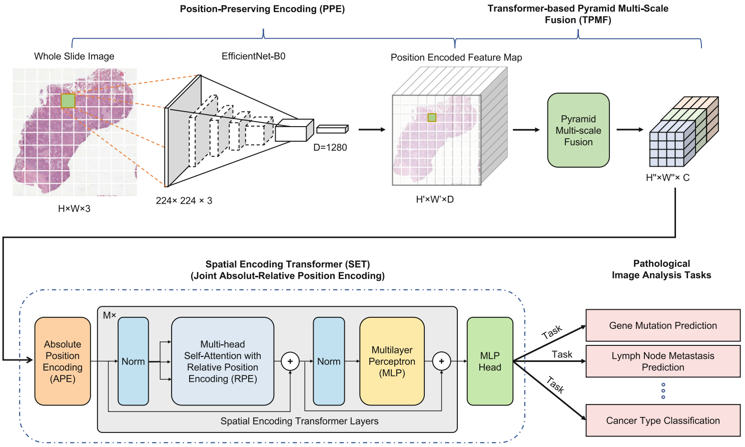
Position-Preserving Encoding (PPE)
Transformer-based Pyramid Multi-Scale Fusion (TPMF)
Whole Slide Image
EfficientNet-B0
Position Encoded Feature Map
H×W×3
224× 224 × 3
D=1280
H'×W'×D
Pyramid Multi-scale Fusion
H''×W''× C
Spatial Encoding Transformer (SET) (Joint Absolut-Relative Position Encoding)
Pathological Image Analysis Tasks
M×
Absolute Position Encoding (APE)
Norm
Multi-head Self-Attention with Relative Position Encoding (RPE)
Norm
Multilayer Perceptron (MLP)
MLP Head
Spatial Encoding Transformer Layers
Task
Task
Task
Gene Mutation Prediction
Lymph Node Metastasis Prediction
Cancer Type Classification

Fig. 1. The overall framework of the proposed spatial encoding transformer-based MIL (SETMIL).

method (SETMIL), which has a spatial encoding transformer-based bag embedding module, joint absolute-relative position encoding mechanism utilizing both absolute position encoding and relative position encoding, and a transformer-based pyramid multi-scale fusion module to comprehensively embed multi-scale information[1] The contributions of this paper include:

1) We present a novel embedded-space MIL method based on the transformer, which has a fully trainable transformer-based bag embedding module to aggregate the bag representation by comprehensively considering all instances.
2) We leverage the spatial encoding transformer to build a MIL method with the advantage of updating the representation of an instance by aggregating representations of both neighbouring instances and globally correlated instances simultaneously, which mimics the clinical practice.
3) We develop a transformer-based pyramid multi-scale fusion module to embed multi-scale information synchronously using multi-scale receptive fields and make the obtained representation enriched with multi-scale context information.
4) We demonstrate that joint absolute-relative position encoding outperforms either of them if utilized independently in transformer-based MIL.

2 Methods

2.1 Overview of the Proposed Method

As illustrated in Fig. 1, the proposed SETMIL consists of three main stages, i.e., position-preserving encoding (PPE), transformer-based pyramid multi-scale

[1] Our code is available at: https://github.com/TencentAILabHealthcare/SETMIL.git.

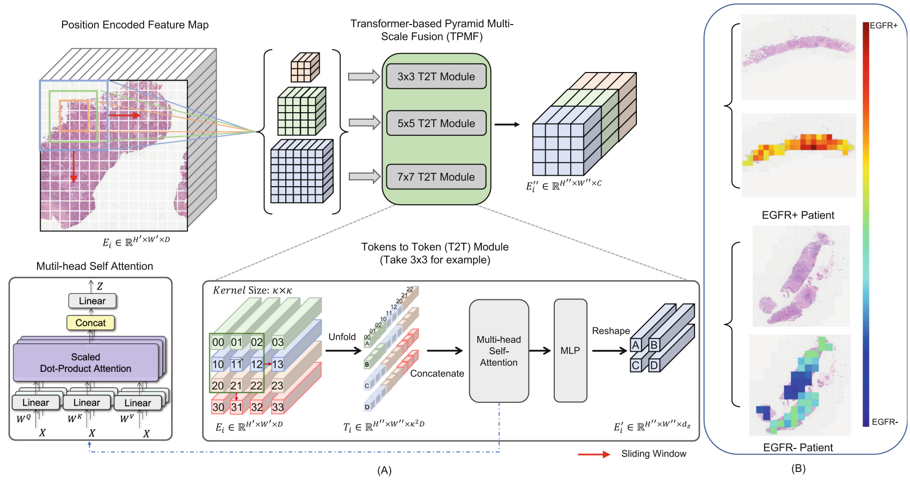
Position Encoded Feature Map
$E_i \in \mathbb{R}^{H' \times W' \times D}$
Transformer-based Pyramid Multi-Scale Fusion (TPMF)
3x3 T2T Module
5x5 T2T Module
7x7 T2T Module
$E_i'' \in \mathbb{R}^{H'' \times W'' \times C}$
Mutil-head Self Attention
Z
Linear
Concat
Scaled Dot-Product Attention
Linear
Linear
Linear
W^Q
W^K
W^V
X
X
X
Tokens to Token (T2T) Module (Take 3x3 for example)
Kernel Size: κ×κ
00 01 02 03
10 11 12 13
20 21 22 23
30 31 32 33
Unfold
Concatenate
Multi-head Self-Attention
MLP
Reshape
A B C D
$E_i \in \mathbb{R}^{H' \times W' \times D}$
$T_i \in \mathbb{R}^{H'' \times W'' \times \kappa^2 D}$
$E_i' \in \mathbb{R}^{H'' \times W'' \times d_z}$
Sliding Window
(A)
EGFR+
EGFR-
EGFR+ Patient
EGFR- Patient
(B)

Fig. 2. Sub-figure (A) illustrates the transformer-based pyramid multi-scale fusion module. Sub-figure (B) shows a example heatmap for model interpretability. Colors reflect the prediction contribution of each local patch.

fusion (TPMF), and spatial encoding transformer (SET)-based bag embedding. First of all, in the PPE stage, SETMIL transforms input huge-size WSI to a small-size position-encoded feature map to simplify the following learning task. Second, the TPMF stage aims at modifying the feature map first and then enables each obtained representation enriched with multi-scale context information. Finally, the SET-based bag embedding module works for generating a high-level bag representation comprehensively considering all instance representations in a fully trainable way and leverages a joint absolute-relative position encoding mechanism to encode the position and context information. The details of these stages are introduced in Sects. 2.2, 2.3, and 2.4 respectively.

2.2 Position-Preserving Encoding

In pathological image analysis, assuming $B_i \in \mathbb{R}^{W \times H \times 3}$ is a WSI, tiled patches from B_i are instances denoted as: $B_i = \{p_i^{0,0}, p_i^{r,c}, \cdots, p_i^{H',W'}\}$, where $p_i^{r,c} \in \mathbb{R}^{\tau_h \times \tau_w \times 3}$, τ_h and τ_w represent the patch size, $H' = H/\tau_h$ and $W' = W/\tau_w$ are the raw and column number of obtained patches. To reduce the scale and facilitate the following learning task, we use a pre-trained EfficientNet-B0 [21] (trained on ImageNet) in the PPE module as the position-preserving encoder $f_E(\cdot)$ to embed each tiled path $p_i^{r,c} \in B_i$ into low-dimensional representations $e_i^{r,c} = f_E(p_i^{r,c}) \in \mathbb{R}^D$. Therefore, the MIL problem can be denoted as:

$$\hat{Y_i} = \phi(g(E_i)) = \phi(g(e_i^{0,0}, e_i^{r,c}, \cdots, e_i^{H',W'})), \tag{1}$$

where $g(\cdot)$ denotes the bag embedding and $\phi(\cdot)$ represents the transformation to predict the label Y_i. After the position-preserving encoding, each WSI is significantly compressed by a factor of $\frac{\tau_h \times \tau_w \times 3}{D}$.

2.3 Transformer-Based Pyramid Multi-scale Fusion

As shown in Fig. 2 (A), the Transformer-based Pyramid Multi-scale Fusion (TPMF) module is composed of three tokens-to-token (T2T) modules [28] working in a pyramid arrangement to modify the feature map and enrich a representation (token) with multi-scale context information. The TPMF is defined as:

$$\begin{aligned} E_i'' &= TPMF(E_i) = Concat(E'_{i,\kappa=3}, E'_{i,\kappa=5}, E'_{i,\kappa=7}) \\ &= Concat(T2T_{\kappa=3}(E_i), T2T_{\kappa=5}(E_i), T2T_{\kappa=7}(E_i)). \end{aligned} \tag{2}$$

The feature modification of TPMF are two-folds, i.e., at one hand can be regarded as the fine-tuning of the obtained feature from the PPE stage and at another hand to reduce the token length as instance-level feature selection. Each T2T module has a softsplit and a reshape process together with a transformer layer [22].

2.4 Spatial Encoding Transformer

The design of SETMIL's spatial encoding transformer (SET) component follows the following principles: (1) generating a high-level bag embedding comprehensively considering the information of all instances in a fully trainable way, (2) updating the representation of an instance by aggregating representations of both neighbour instances and globally correlated instances simultaneously mimicking the clinical practice, and (3) jointly leveraging absolute and relative position encoding to strengthen the context information encoding ability of SETMIL. As shown in Fig. 1, the SET has an absolute position encoding process at the beginning, M stacked spatial-encoding transformer layers (SETLs, $M = 6$) that jointly consider global correlation and local context information similarity to generate the bag embedding (as $g(\cdot)$) using relative spatial encoding, and a MLP to map the bag-embedding into the prediction (as $\phi(\cdot)$). We utilize the sinusoid procedure in standard transformer [22] as the absolute position encoding method. Besides, similar to [24,27], we apply the layer normalization (LN) before instead of after multi-head self-attention (MSA) operation and a multilayer perceptron (MLP) in each spatial encoding transformer layer. The prediction of SETMIL can be denoted as:

$$\hat{Y_i} = \phi(g(E_i'')) = \phi(SETL(\underset{M=6}{\cdots}, SETL(E_i''))), \tag{3}$$

where $SETL(\cdot)$ is defined as:

$$SETL(E_i'') = MLP(MSA^{SET}(LN(E_i''))). \tag{4}$$

The MSA^{SET} represents the multi-head self-attention in SETL, which is similar as MSA in [22] but with relative position encoding (Sect. 2.4) to embed the position and context information of each WSI. The self-attention mechanism in MSA is typical key-value attention. Assuming there is an input sequence $X = (x_1, x_2, \cdots, x_N)$, the output sequence of the self attention $Z = (z_1, z_2, \cdots, z_N)$ can be calculated as:

$$z_i = \mathcal{A}(X, W^Q, W^K, W^V) = \sum_{j=1}^{N} \frac{exp(\eta_{ij})}{\sum_{k=1}^{N} exp(\eta_{ik})}(x_j W^V), \tag{5}$$

where $x_i \in \mathbb{R}^{d_x}$, $z_i \in \mathbb{R}^{d_z}$, $W^Q, W^K, W^V \in \mathbb{R}^{d_x \times d_z}$ are the parameter projection matrices, η_{ij} is computed by a scaled dot-product attention:

$$\eta_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}}. \tag{6}$$

Relative Position Encoding. The mechanism of the $MSA^{SET}(\cdot)$ is demonstrated in Fig. 3. For two instance representations e''_i and e''_j in feature map $E'' = (e''_1, e''_2, \cdots, e''_{H''W''})$ with 2D coordinate of $\rho_i = (r''_i, c''_i)$. The Euclidean distance μ between two instances can be calculated as:

$$\mu(\rho_i, \rho_j) = \sqrt{(r''_i - r''_j)^2 + (c''_i - c''_j)^2}. \tag{7}$$

To encode the position information, we refer the relative position encoding idea and assign a learnable scalar serving as a bias term in the self-attention module in formula (6). Therefor we have:

$$\eta''_{i,j} = \frac{(e''_i W^Q)(e''_j W^K)^T}{\sqrt{d''_z}} + \lambda_{\mu(\rho_i, \rho_j)}, \tag{8}$$

where d''_z is the dimension of SETL's output z''_i and $\sqrt{d''_z}$ is used for appropriate normalization. $\lambda_{\mu(\rho_i, \rho_j)}$ is the learnable scalar indexed by $\mu(\rho_i, \rho_j)$, which is defined as: $\lambda_{\mu(\rho_i, \rho_j)} = \theta(\mu(\rho_i, \rho_j))$, where $\theta(\cdot)$ is learnable to adaptively assign weights for different spatial distances. As shown in Fig. 3, the first item in formula (8) is the same as a conventional transformer, which represents updating the representation of an instance referring to correlations between current instance and other instances globally. While, on the other hand, the second item assigns the same weight for instances with the same spatial distance to the current instance. This spatial-distance aware encoding strategy is similar to CNN but its kernel size is adaptively adjusted by $\theta(\cdot)$ and therefore has to potential to update the representation of an instance by aggregating neighbouring instances.

3 Experiments

Dataset: (1) Gene Mutation Prediction: The first task to evaluate the performance of SETMIL is gene mutation (GM) prediction. A totol of 723 WSI slides

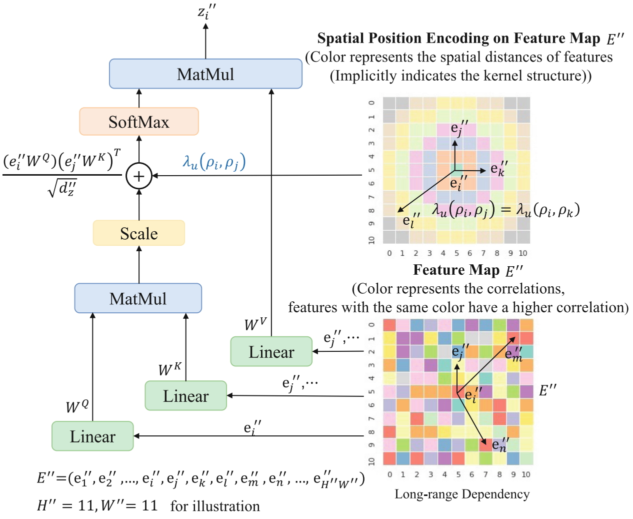
z_i''
MatMul
SoftMax
$\frac{(e_i''W^Q)(e_j''W^K)^T}{\sqrt{d_z''}}$
$\lambda_u(\rho_i,\rho_j)$
Scale
MatMul
W^V
Linear
W^K
Linear
W^Q
Linear
$e_j'',\cdots$
$e_j'',\cdots$
e_i''
Spatial Position Encoding on Feature Map E''
(Color represents the spatial distances of features
(Implicitly indicates the kernel structure))
e_j''
e_k''
e_i''
e_l''
$\lambda_u(\rho_i,\rho_j)=\lambda_u(\rho_i,\rho_k)$
0 1 2 3 4 5 6 7 8 9 10
Feature Map E''
(Color represents the correlations,
features with the same color have a higher correlation)
e_j''
e_m''
e_i''
e_n''
E''
0 1 2 3 4 5 6 7 8 9 10
Long-range Dependency
$E''=(e_1'',e_2'',\dots,e_i'',e_j'',e_k'',e_l'',e_m'',e_n'',\dots,e_{H''W''}'')$
$H''=11, W''=11$ for illustration

Fig. 3. Illustration of the spatial encoding transformer layer.

from patients diagnosed with LUAD were collected, where 47% cases included are with the epidermal growth factor receptor (EGFR) gene mutation. In this task, we use SETMIL to predict whether a patient is with EGFR gene mutation. (2) Lymph Node Metastasis Diagnosis: Another challenging task we would like to solve with SETMIL is the prediction of lymph node metastasis (LNM) status. 1274 WSI slides from patients with endometrial cancer (EC) were collected, among which 44% cases are diagnosed with lymph node metastasis. We apply this dataset to assess the potential of SETMIL in the prediction of LNM status.

Implementation Details: The samples in every dataset are randomly divided into the training, validation and test set with the percentages of 60%, 20%, 20%. We tile the WSI into patches using a 1120 × 1120 pixels sliding window without overlap. The proposed model is implemented in Pytorch and trained on one 32 GB TESLA V100 GPU. We utilize the Cross-Entropy Loss and the AdamW optimizer [12] with a learning rate of $2e^{-4}$ and a weight decay of $1e^{-4}$. The batch size is set to 4.

4 Results and Discussion

Comparison with State-of-the-Art Methods: The performance of SETMIL and other state-of-the-art (SOTA) methods including ABMIL [7,13], DSMIL [10], CLAM [14], RNN-MIL [2], ViT-MIL [4], TransMIL [17] and CNN-MIL are compared in Table 1. All methods are evaluated in two tasks, i.e., GM prediction (with/without EGFR) and LNM prediction (with/without LNM). Generally, the SETMIL achieves 83.84% AUC in the GM prediction task and 96.34% AUC in the LNM prediction task. From Table 1, we can also find that SETMIL outperforms other SOTA methods in the two tasks with over 1.51% and 5.19% performance enhancement (AUC), respectively. Figure 2 (B) shows a sample heatmap of the SETMIL reflecting the prediction contribution of each local patch. The top contributed patches for the prediction of EGFR mutation positive (EGFR+) and EGFR mutation negative (EGFR-) are indicated with red and blue colours, respectively.

Table 1. The performance of SETMIL compared with other state-of-the-art methods.

	LUAD-GM					EC-LNM				
Models	AUC (%)	Accuracy	Precision	Recall	F1-score	AUC (%)	Accuracy	Precision	Recall	F1-score
ABMIL [7,13]	52.44	54.55	47.83	13.10	20.56	64.30	53.33	53.33	**99.99**	69.57
RNN-MIL [2]	50.31	56.68	53.49	27.38	36.22	86.89	81.18	74.18	99.26	84.91
CNN-MIL	45.28	44.92	44.92	**99.99**	61.99	88.06	81.96	75.57	97.79	85.26
DSMIL [10]	78.53	71.53	**89.19**	47.14	61.68	88.51	85.10	79.88	96.32	87.33
CLAM-SB [14]	78.49	70.14	70.15	67.14	68.91	86.96	78.93	82.49	78.49	80.44
CLAM-MB [14]	82.33	75.70	73.97	77.14	75.52	87.62	81.31	83.24	82.80	83.02
ViT-MIL [4]	76.39	70.14	81.40	50.00	61.95	91.15	86.27	80.24	98.53	88.45
TransMIL [17]	77.29	74.45	68.74	69.02	68.87	82.76	74.78	74.58	73.98	74.16
SETMIL(Ours)	**83.84**	**76.38**	71.14	86.08	**78.24**	**96.34**	**92.94**	**92.75**	94.12	**93.43**

Ablation Studies: (1) Effects of Proposed Components: To evaluate the effectiveness of each proposed component in the SETMIL, we conducted experiments on the following configurations: (A) SETMIL (our proposed method): PPE + TPMF (T2T module & Pyramid Multi-scale Fusion (PMF) idea) + SET; (B) “w/o TPMF&SET”: SETMIL without using both the TPMF and SET module, which is the same as the ViT-MIL [4]; (C) “w/o PMF&SET” SETMIL without using both the pyramid multi-scale fusion (PMF) idea and SET module but having a single 5×5 T2T module for feature modification and dimension reduction; (D) “w/o SET”: SETMIL without using SET; (E) “w/o PMF”: SETMIL without using PMF but having a single 5×5 T2T module for feature modification and dimension reduction; The ablation study results are illustrated in Table 2, where we can conclude that the SETMIL benefits from each proposed components. (2) Assessment of Different Position Encoding Strategies: To assess the performance of SETMIL by using different position encoding strategies, we also conducted experiments on the following two configurations: (A) “w/o Absolute”: SETMIL without using absolute position encoding and (B) “w/o Relative”: SETMIL without using relative position encoding (same as “w/o SET”). The experiment results are shown in Table 2, indicating that using both absolute position encoding and relative position encoding mechanisms jointly improves the performance compared to using either one of them.

Table 2. Ablation studies: effects of proposed components.

	LUAD-GM					EC-LNM				
Methods	AUC (%)	Accuracy	Precision	Recall	F1-score	AUC (%)	Accuracy	Precision	Recall	F1-score
w/o TPMF&SET	76.39 (−7.45)	70.14	81.40	50.00	61.95	91.15 (−5.19)	86.27	80.24	**98.53**	88.45
w/o PMF&SET	79.40 (−4.44)	70.83	65.56	84.29	**73.75**	92.84 (−3.50)	87.06	85.03	91.91	88.34
w/o SET	81.27 (−2.57)	72.92	78.18	61.43	68.80	95.42 (−0.92)	91.76	89.66	95.59	92.53
w/o PMF	80.41 (−3.43)	61.11	56.36	**88.57**	68.89	95.85 (−0.49)	89.41	87.59	93.38	90.39
w/o Absolute	75.56 (−8.28)	72.22	70.27	**74.29**	**72.22**	89.32 (−7.02)	80.39	85.25	76.47	80.62
w/o Relative	81.27 (−2.57)	72.92	78.18	61.43	68.80	95.42 (−0.92)	91.76	89.66	**95.59**	92.53
SETMIL(Ours)	**83.84**	**76.38**	71.14	86.08	**78.24**	**96.34**	**92.94**	**92.75**	94.12	**93.43**

5 Conclusion

In this paper, we comprehensively considered the characteristics of pathological image analysis and presented a novel spatial encoding transformer-based MIL method, which has the potential to be a backbone for solving challenging pathological image analysis tasks. Experimental results demonstrated the superior performance of the proposed SETMIL compared to other state-of-the-art methods. Ablation studies also demonstrated the contribution of each proposed component of SETMIL.

References

1. Amores, J.: Multiple instance classification: review, taxonomy and comparative study. Artif. Intell. **201**, 81–105 (2013)
2. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (2019)
3. Diao, J.A., et al.: Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nat. Commun. **12**(1), 1–15 (2021)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
5. Garrett, W.S.: Cancer and the microbiota. Science **348**(6230), 80–86 (2015)
6. Hashimoto, N., et al.: Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3852–3861 (2020)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
8. Kandemir, M., Hamprecht, F.A.: Computer-aided diagnosis from weak supervision: a benchmarking study. Computeriz. Med. Imaging Graph. **42**, 44–50 (2015)
9. Kather, J.N., et al.: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. **25**(7), 1054–1056 (2019)

10. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2021)
11. Li, R., Yao, J., Zhu, X., Li, Y., Huang, J.: Graph CNN for survival analysis on whole slide pathological images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 174–182. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_20
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
13. Lu, M.Y., et al.: Ai-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (2021)
14. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**(6), 555–570 (2021)
15. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 893–901. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_99
16. Rubin, R., et al.: Rubin's Pathology: Clinicopathologic Foundations of Medicine. Lippincott Williams & Wilkins (2008)
17. Shao, Z., et al.: Transmil: transformer based correlated multiple instance learning for whole slide image classication. arXiv preprint arXiv:2106.00908 (2021)
18. Skrede, O.J., et al.: Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet **395**(10221), 350–360 (2020)
19. Song, Z., et al.: Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat. Commun. **11**(1), 1–9 (2020)
20. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: a survey. Med. Image Anal. 101813 (2020)
21. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
22. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
23. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recogn. **74**, 15–24 (2018)
24. Xiong, R., et al.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning, pp. 10524–10533. PMLR (2020)
25. Yan, Y., Wang, X., Guo, X., Fang, J., Liu, W., Huang, J.: Deep multi-instance learning with dynamic pooling. In: Asian Conference on Machine Learning, pp. 662–677. PMLR (2018)
26. Yao, J., Zhu, X., Huang, J.: Deep multi-instance learning for survival prediction from whole slide images. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 496–504. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_55
27. Ying, C., et al.: Do transformers really perform bad for graph representation? arXiv preprint arXiv:2106.05234 (2021)
28. Yuan, L., et al.: Tokens-to-token vit: training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986 (2021)

29. Zhou, Y., Onder, O.F., Dou, Q., Tsougenis, E., Chen, H., Heng, P.-A.: CIA-Net: robust nuclei instance segmentation with contour-aware information aggregation. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 682–693. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_53
30. Zhou, Y., Sun, X., Liu, D., Zha, Z., Zeng, W.: Adaptive pooling in multi-instance learning for web video annotation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 318–327 (2017)