Tumor Microenvironment-Guided Fine-Tuning of Pathology Foundation Models for Esophageal Squamous Cell Carcinoma Immunotherapy Response Prediction

Yixuan Lin[1], Weiping Lin[1], Chenxu Guo[2], Xinxin Yang[2], Hongxue Meng[2(✉)], and Liansheng Wang[1(✉)]

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
{yixuanlin, wplin}@stu.xmu.edu.cn, lswang@xmu.edu.cn

[2] Department of Pathology, Harbin Medical University Cancer Hospital, Harbin, China
{guochenxu0915, 13946915562, menghongxue15}@163.com

Abstract. Esophageal squamous cell carcinoma (ESCC) has high incidence and mortality rates. While immunotherapy shows promise for some ESCC patients, others can experience severe side effects. Accurate pre-screening of individual patients' immunotherapy response to ESCC is a crucial but difficult task. Subtle differences in pre-treatment biomarkers hinder physicians' judgment in pathological diagnosis. While pathological foundation models (PFMs) have shown potential in pathology image analysis, traditional PFMs focused on image-level features still struggle to capture nuanced preoperative characteristic differences. To address this, we propose a fine-tuning framework for PFMs based on the tumor microenvironment (TME). First, morphological and topological attributes are extracted from larger field-of-view patches to better analyze TME interactions. Next, we utilize PFMs which are typically constrained to small inputs to extract image features. To address this limitation, larger patches are subdivided to prevent precision loss, with trainable position encodings maintaining relative spatial positional relationships to guide the re-aggregation of large patch-level representations. Finally, a TME-guided learning algorithm trains all trainable layers to understand ESCC-specific characteristics. Our framework demonstrates superior performance in the downstream task of predicting ESCC immunotherapy response compared to those fine-tuned using self-supervised learning methods. By allowing flexibility in patch sizes, our approach captures more contextual information. Code is available at https://github.com/stoney03/ESCC.

Keywords: Immunotherapy · Tumor microenvironment · Position encodings · Histopathology.

Y. Lin and W. Lin contributed equally.

1 Introduction

Esophageal cancer is the 11th most common cancer globally and the 7th leading cause of cancer-related deaths worldwide [2]. Esophageal squamous cell carcinoma (ESCC) constitutes approximately 90% of annual diagnosed esophageal cancers [1,15]. Due to the lack of specific early symptoms, ESCC remains a major global health issue with a large clinical burden [21].

Immunotherapy has improved the treatment of ESCC [13]. However, some patients experience severe adverse effects including immune-related inflammation and organ toxicity [11]. Thus, physicians perform preoperative screening using histopathology [16] to determine treatment efficacy. Unfortunately, the preoperative differences between treatment responders and non-responders are often subtle and physicians struggle to predict future clinical outcomes.

Deep learning techniques offer powerful tools to analyze whole slide images (WSI), achieving performance that surpasses pathologists in specific diagnostic tasks [3,5]. Most WSI analysis combines pathological foundation models (PFMs) with multiple instance learning (MIL) methods [8,14,12,17,6]. Unfortunately, these approaches have struggled in predicting clinical outcomes of ESCC immunotherapy [18]. This may be attributed to PFMs' focus on image-level features which, fail to capture more subtle characteristics. In contrast, cellular interactions within the tumor microenvironment (TME) constitute more complex and meaningful features [4,22]. TME-guided learning algorithms offer a promising solution: learning tumor-related information from larger field-of-view patch-level images, guiding the PFM model to extract beyond image-level features and instead interpret TME characteristics, then generalizing to downstream tasks.

In this study, we present a novel TME-guided PFM fine-tuning framework designed to predict ESCC immunotherapy efficacy. Recognizing that small patches may hinder information acquisition from contextual interactions within the TME, we employ larger patches to capture multi-cellular interactions. Next, we use a PFM to extract image features. As most PFMs are designed for small-size patches, we divide large patches into sub-patches to avoid compression loss. We introduce adaptive position encodings to preserve relative spatial relationships among embeddings critical for TME analysis. We also add a learnable feature aggregation module to reconstruct the original larger-patch representations. The TME information is subsequently used to fine-tune all trainable layers (i.e. unfrozen layers in PFM, position encoding, aggregation model) with a better understanding of preoperative ESCC-specific differences between immunotherapy responders and non-responders. In summary, the primary contributions of our work are listed as follows:

(1) We introduce a novel, annotation-free fine-tuning framework for PFMs using the TME to enhance the understanding of latent feature within WSI. Our work improves model performance in downstream tasks for predicting ESCC immunotherapy response.

(2) Our framework supports WSI analysis with flexibility in patch size while minimizing contextual information loss. Large patches are subdivided down into sub-patches that maintain relative spatial positional relationships and are then

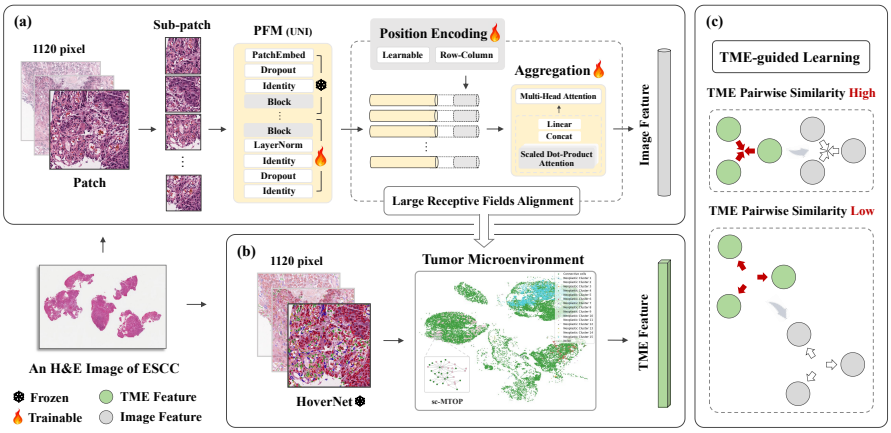

Fig. 1: Overview of our TME-guided fine-tuning framework for PFMs, with more flexibility in patch size.

re-aggregated into a single informative representation. This method is simple, effective, and easy to transfer to other pathology tasks.

(3) Our framework significantly improves the recall and F1 score of patients who will respond to ESCC immunotherapy, as demonstrated by quantitative evaluations on our collected dataset. This improvement has important clinical implications.

2 Methodology

2.1 Overview

An overview of our framework to improve TME representation within PFMs is shown in Fig. 1. This work begins by processing sub-patches derived from large-scale patches through a PFM, where trainable positional encodings guide the spatial re-aggregation of features into unified large-patch representations. Subsequently, morphological and topological attributes are extracted from large patches to construct TME representations. Finally, we use the similarity patterns between TME features to guide and adjust the corresponding image patch features, fine-tuning all trainable layers in the model. This approach leads to a better performance in predicting ESCC future treatment outcomes.

2.2 Tumor Microenvironment Information Extraction

As illustrated in Fig. 1(b), we incorporate TME information based on clinical experts' prior knowledge [9,11] to bridge the semantic gap between complex visual patterns and their diagnostic significance in data-limited scenarios. First, we analyze the morphological characteristics and proportional distribution of

various cell types within the TME, with particular focus on the opposing effects of fibroblasts and inflammatory cells on tumor cells. Since tumor treatment depends on their interaction with other cellular components, it is necessary to construct an atlas to explore tumor-related ecosystem features. Therefore, we use a pretrained HoverNet [7] model to perform patch-level segmentation and classification of cells in WSIs, then we quantify their spatial distribution and relative abundance. Finally, we utilize the sc-MTOP [20] method at the patch level to characterize the ecosystem among tumor cells, inflammatory cells, and stromal cells. This approach is consistent with the notion that spatially resolved analysis can identify recurrent micro-ecological modules.

2.3 Large Receptive Fields Alignment

TME features containing rich spatial-level contextual information are obtained at the large patch level. However, most PFMs can only intake small-sized patches. Therefore, after subdividing large patches into sub-patches (Fig. 1(a)), we use position encodings to maintain the relative spatial position information of the sub-patches during the re-aggregation process.

Position Encoding is composed of two parts: the **learnable position encoding** and the **row-column encoding**. A large patch is subdivided into a set of discretized sub-patches $\{p_1, p_2, \ldots, p_N\}$, where N is the total number of sub-patches in the large patch. Each sub-patch p_i is mapped to a vector $\mathbf{e}_i \in \mathbb{R}^d$ in the embedding space using a PFM, where d is the embedding dimension.

(a) Learnable Position Encoding: This position encoding is represented by a learnable parameter matrix $\mathbf{P} \in \mathbb{R}^{1\times N\times d}$, which is initialized with a standard normal distribution. This matrix is updated during training through gradient descent and aims to provide a unique spatial representation for each patch.

(b) Row-Column Encoding: To incorporate the spatial position (row and column) of each patch, we first define the row and column indices: $r_i, c_i \in \mathbb{R}$ normalized to the range [0,1]. To encode the row and column information into the embedding space, we construct a new matrix $\mathbf{R} \in \mathbb{R}^{1\times N\times d}$ as the row-column encoding, where the final dimension stores normalized row positions in the first $\frac{d}{2}$ entries and normalized column positions in the second $\frac{d}{2}$ entries. Thus, the row-column encoding $\mathbf{R}$ is represented as:

$$\mathbf{R}_{i,:d/2} = \frac{r_i}{H-1}, \quad \mathbf{R}_{i,d/2:} = \frac{c_i}{W-1}, \tag{1}$$

where H and W are the number of rows and columns of sub-patches respectively. The final position encoding $\mathbf{PE} \in \mathbb{R}^{1\times N\times d}$ is the sum of two components above. The input feature map $\mathbf{x} \in \mathbb{R}^{N\times d}$ is then augmented with $\mathbf{PE}$ to form the final input representation.

Sub-patch Aggregation We use a trainable combination of multi-head self-attention (MSA) and a convolutional layer to aggregate features from sub-patches. We first apply MSA to input features $\mathbf{Z} \in \mathbb{R}^{N\times d}$, then the attention

weights $\mathbf{A} \in \mathbb{R}^{1\times N\times N}$ are used to compute a weighted sum of the attention output $\mathbf{O} \in \mathbb{R}^{1\times N\times d}$, pooling the patch features: $\mathbf{f}_{attn\_pool} = Squeeze(\mathbf{A}\cdot\mathbf{O}^T) \in \mathbb{R}^{N\times d}$. Next, a convolutional layer $Conv(\cdot)$ is applied to the aggregated features to reduce the dimensionality and captures spatial patterns by the following formula:

$$\mathbf{f}_{agg} = \frac{1}{N}\sum_{i=1}^{N} Conv\ (\mathbf{f}_{attn\_pool}). \tag{2}$$

2.4 TME-guided Learning

We propose an algorithm using TME information to fine-tune whole trainable layers, focusing on ESCC-specific features (Fig. 1(c)). We define a batch of aggregated image features as $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M\}$ and TME features as $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M\}$, where M is the size of a batch, and each $\mathbf{f}_i \in \mathbb{R}^d$ and $\mathbf{c}_i \in \mathbb{R}^m$ represents the i-th patch respectively. The pairwise similarity between $\mathbf{f}_i$ and $\mathbf{f}_j$ is calculated using the cosine similarity, and the same as $\mathbf{c}_i$ and $\mathbf{c}_j$. Due to the symmetric nature of the pairwise similarity matrix, we optimize computations by calculating only the upper triangular portion of the matrices. The TME-guided loss function aims to minimize the discrepancy between image feature and TME feature:

$$\mathcal{L}_{TG} = \frac{1}{\binom{m}{2}} \sum_{1\le i<M} \sum_{i<j\le M} \left(S_t(i,j) - S_i(i,j) - margin\right)^2, \tag{3}$$

where $S_t(i,j)$ and $S_i(i,j)$ represent TME and image feature similarity, respectively.

3 Experiments

Datasets Our dataset consists of 128 ESCC patients' pre-treatment H&E WSIs, including 55 responders and 73 non-responders to immunotherapy. We identified WSIs with consecutive slices and only retained unique slices to both streamline the subsequent processing and mitigate the risk of overfitting. We adopted a five-fold cross-validation approach (aggregate predicted results to calculate the overall evaluation metric), then split the dataset into training, validation, and testing sets randomly with a proportion of 6:2:2 for comprehensive evaluation.

Table 1: Pre-experiment with different feature types. ❄: frozen; &: concatenation.

Feat Type	**Method**	**ACC**	**AUC**	**Recall$_R$**	**F1$_R$**
UNI[5]❄	CLAM[14]	0.5469	0.5756	0.1636	0.2368
TME-W	MLP	0.6328	0.5768	0.5455	0.5607
UNI[5]❄ & TME-W	CLAM[14]	0.5703	0.5755	0.3455	0.4086
Ours	CLAM[14]	**0.7094**	**0.7274**	**0.6154**	**0.6273**

Table 2: Five-fold cross-validation performance of ESCC immunotherapy prediction on various downstream models (DM).

DM		UNI[5]				Prov-GigaPath[19]			
		ACC	**AUC**	**Recall**$_R$	**F1**$_R$	**ACC**	**AUC**	**Recall**$_R$	**F1**$_R$
ABMIL[8]	Baseline	0.6172	0.5589	0.3455	0.3838	0.6094	0.6087	0.3818	0.4565
	Ours	0.6703	0.6771	0.5455	0.6000	0.7016	0.6915	0.5909	0.5821
	Δ	0.0531	0.1182	0.2000	0.2162	0.0922	0.0828	0.2091	0.1256
CLAM[14]	Baseline	0.6172	0.6481	0.3636	0.4494	0.5938	0.5636	0.3455	0.4176
	Ours	**0.7094**	**0.7274**	**0.6154**	**0.6273**	0.6938	0.6579	0.5727	**0.6000**
	Δ	0.0922	0.0793	0.2518	0.1779	0.1000	0.0943	0.2272	0.1824
DSMIL[12]	Baseline	0.6094	0.6192	0.3091	0.4048	0.6094	0.5793	0.3273	0.4186
	Ours	0.6859	0.7167	0.5391	0.5636	**0.7172**	**0.6963**	0.6000	0.5739
	Δ	0.0765	0.0975	0.2300	0.1588	0.1078	0.1170	0.2727	0.1553
TransMIL[17]	Baseline	0.5625	0.5953	0.3818	0.4286	0.5703	0.5469	0.3636	0.4211
	Ours	0.6625	0.6980	0.5484	0.6182	0.6594	0.6943	**0.6364**	0.5833
	Δ	0.1000	0.1027	0.1666	0.1896	0.0891	0.1474	0.2728	0.1622

Implementation Details All WSIs were tiled into non-overlapping patches of 1120 × 1120 pixels at 40× magnification. The first freezing layers of PFMs (UNI [5], Prov-GigaPath [19]) were set to 300. These patches were further divided into sub-patches of 224 × 224 pixels to match PFMs' input size, yielding final feature dimensions 1 × 1024 and 1 × 1536, respectively. The experiment was conducted on a GeForce RTX 3090. For fine-tuning, we trained the layers for 20 epochs using the Adam optimizer [10]. We used a learning rate of 0.0001 and a batch size of 4 with gradient accumulation and mixed precision training. A 50-epoch training protocol was adopted for MIL methods with a batch size 16.

Preliminary Experiment Our approach leverages PFMs' pretrained weights and TME features, prompting pre-experiments to assess their performance. We evaluated five-fold cross-validation results with frozen UNI image features (224 pixels), TME features in WSI level (TME-W), and their concatenation. TME-W were classified by Multilayer Perceptron (MLP), while others used the MIL method CLAM [14]. Table 1 indicates that TME contributed to performance improvement, prompting the use of it to fine-tune PFMs in our framework.

Baseline Comparison We use accuracy (**ACC**), the area under the receiver operating characteristic curve (**AUC**), Recall (**Recall**$_R$), and F1 score (**F1**$_R$) as metrics for evaluation, where R stands for the class *response to immunotherapy*. For our experiments, we fine-tuned unfrozen layers of PFMs (UNI and Prov-GigaPath), position encoding, and aggregation module. For a baseline, we use the same unfrozen layers of PFMs fine-tuned using the self-supervised learning approach, with input size 224 pixels. We then conducted a comprehensive evaluation using classical and advanced WSI analysis methods (ABMIL [8], CLAM,

Table 3: Performance evaluation of our framework with different input patch sizes using UNI with CLAM.

Patch Size	ACC	AUC	Recall_R	F1_R
672 (3×224) px	0.6781	0.6851	0.5818	0.5424
896 (4×224) px	0.6859	0.6888	0.5893	0.6225
1120 (5×224) px	**0.7094**	**0.7274**	**0.6154**	**0.6273**
1344 (6×224) px	0.6759	0.7012	0.6000	0.5982

DSMIL [12] and TransMIL [17]) with a five-fold cross-validation strategy to assess the performance of both the baseline and our proposed fine-tuning framework. Table 2 shows that our framework significantly outperforms the baseline across multiple MIL methods for ESCC immunotherapy respond prediction. Specifically, we achieved a remarkable 63.64% Recall_R and 71.72% ACC using Prov-GigaPath, demonstrating substantial gains over the baseline. The high recall for the benefit class is particularly promising for clinical applications.

Table 4: Ablation Study of key components of our framework using UNI and CLAM (input size: 1120 px). Agg: aggregation model; PE: row-column encoding.

Method		ACC	AUC	Recall_R	F1_R
w/o TME	UNI[5]+Agg+SSL	0.6250	0.6286	0.4000	0.4783
	UNI[5]+PE+Agg+SSL	0.6328	0.6077	0.4000	0.4835
w/o Learnable PFM	UNI[5]❄+Agg+TME	0.6484	0.6262	0.4364	0.5161
	UNI[5]❄+PE+Agg+TME	0.6562	0.6174	0.4545	0.5319
w/o Conv layer	UNI[5]+MSA+TME	0.6250	0.6618	0.5455	0.5556
	UNI[5]+PE+MSA+TME	0.6875	0.7004	0.6000	0.6226
w/o PE	UNI[5]+Agg+TME	0.6679	0.6783	0.5818	0.5614
Ours	UNI[5]+PE+Agg+TME	**0.7094**	**0.7274**	**0.6154**	**0.6273**

Ablation Study We conducted a series of ablation studies to validate our framework's key steps. (1) *Large Receptive Fields*: Results in Table 3 suggest a larger field of vision helps PFMs better comprehend information in the TME. However, accuracy drops as input size reaches 1344 pixels, likely due to challenges in learning contextual relationships in images that are too large. (2) *Tumor Microenvironment Features*: We removed patch-level TME features and used self-supervised learning to fine-tune UNI. Table 4 **w/o TME** shows how the base model is not suitable for ESCC-specific tasks. (3) *PFM Freezer*: We froze UNI's layers to isolate its impact on the experimental results. Table 4 **w/o Learnable PFM** reveals that the incorporation of TME features is crucial for the effective

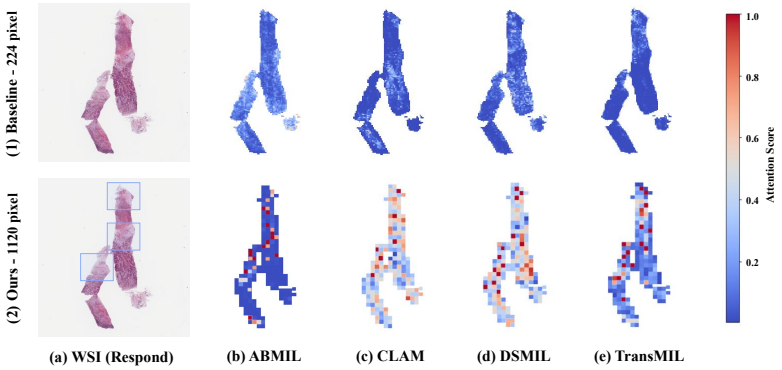

Fig. 2: Attention heatmaps comparison for WSI analysis methods: ABMIL, CLAM, DSMIL and TransMIL. The blue boxes indicate the main areas of tumor tissues with abundant stroma.

fine-tuning of PFMs, as freezing the UNI model resulted in minimal improvements in prediction. (4) *Convolutional Layer*: We demonstrated that removing the convolutional layer from the aggregation module led to a decrease in performance (Table 4 **w/o Conv layer**), showing its necessity in complementing the multi-head self-attention mechanism and improving feature representation. (5) *Row-Column Encoding*: We aggregated sub-patch features without row-column encoding and found 6.59% decrease in $F1_R$ (Table 4 **w/o PE**). This demonstrates that positional encodings are effective at aligning with TME spatial distribution. The results of our ablation studies show that all modules in our approach are essential, each improving the PFM's performance.

Attention Map Visualization Fig. 2 illustrates an H&E image belonging to an immunotherapy responder with attention heatmaps generated by four WSI analysis methods: ABMIL, CLAM, DSMIL, and TransMIL. The features utilized for Fig. 2(1) and Fig. 2(2) are extracted by UNI after fine-tuning with the baseline method and our proposed framework, respectively. In the first row, we observe that all models exhibit nearly uniform attention across the WSI, with no distinct regions of focus. In contrast, features from the TME-guided fine-tuning framework in the second row help downstream models to generate distinct regions of attention that focus on tumor tissues with abundant stroma, which are associated with immunotherapy.

4 Conclusion

We propose a novel TME-guided fine-tuning framework for PFMs to predict ESCC immunotherapy efficacy using histological WSIs. Unlike existing methods, our framework captures subtle biomarker variations using patch subdivi-

sion to extract diverse TME features across the WSI; trainable position encodings to effectively model spatial positional relationships within patches; and a TME-guided learning algorithm to teach models richer and more robust representations. Our findings demonstrate superior performance compared to self-supervised learning methods, highlighting the effectiveness of our proposed approach. This will pave the way for more personalized and effective immunotherapy treatment for ESCC patients.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abnet, C.C., Arnold, M., Wei, W.Q.: Epidemiology of esophageal squamous cell carcinoma. Gastroenterology **154**(2), 360–373 (2018)
2. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)
3. Bulten, W., Pinckaers, H., Boven, H.V., Vink, R., Litjens, G.: Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology **21**(2), 233–241 (2020)
4. Chang, C.H., Qiu, J., O'Sullivan, D., Buck, M., Noguchi, T., Curtis, J., Chen, Q., Gindin, M., Gubin, M., Vanderwindt, G.W.: Metabolic competition in the tumor microenvironment is a driver of cancer progression. Cell **162**(6), 1229–1241 (2015)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
6. Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.H., Christodoulidis, S.: Structured state space models for multiple instance learning in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–604. Springer (2023)
7. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis **58**, 101563 (2019)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. Kalluri, R., Zeisberg, M.: Fibroblasts in cancer. Nature reviews cancer **6**(5), 392–401 (2006)
10. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kojima, T., Doi, T.: Immunotherapy for esophageal squamous cell carcinoma. Current oncology reports **19**, 1–8 (2017)

12. 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)
13. Li, N., Sohal, D.: Current state of the art: immunotherapy in esophageal cancer and gastroesophageal junction cancer. Cancer Immunology, Immunotherapy **72**(12), 3939–3952 (2023)
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. Nature biomedical engineering **5**(6), 555–570 (2021)
15. Morgan, E., Soerjomataram, I., Rumgay, H., Coleman, H.G., Thrift, A.P., Vignat, J., Laversanne, M., Ferlay, J., Arnold, M.: The global landscape of esophageal squamous cell carcinoma and esophageal adenocarcinoma incidence and mortality in 2020 and projections to 2040: new estimates from globocan 2020. Gastroenterology **163**(3), 649–658 (2022)
16. Qu, H.T., Li, Q., Hao, L., Ni, Y.J., Luan, W.Y., Yang, Z., Chen, X.D., Zhang, T.T., Miao, Y.D., Zhang, F.: Esophageal cancer screening, early detection and treatment: Current insights and future directions. World Journal of Gastrointestinal Oncology **16**(4), 1180 (2024)
17. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
18. Wang, X., Barrera, C., Bera, K., Viswanathan, V.S., Azarianpour-Esfahani, S., Koyuncu, C., Velu, P., Feldman, M.D., Yang, M., Fu, P., et al.: Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (tils) predict clinical benefit for immune checkpoint inhibitors. Science advances **8**(22), eabn3966 (2022)
19. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
20. Zhao, S., Chen, D.P., Fu, T., Yang, J.C., Ma, D., Zhu, X.Z., Wang, X.X., Jiao, Y.P., Jin, X., Xiao, Y., et al.: Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. Nature Communications **14**(1), 6796 (2023)
21. Zhao, Y.X., Zhao, H.P., Zhao, M.Y., Yu, Y., Qi, X., Wang, J.H., Lv, J.: Latest insights into the global epidemiological features, screening, early diagnosis and prognosis prediction of esophageal squamous cell carcinoma. World Journal of Gastroenterology **30**(20), 2638 (2024)
22. Zou, W.: Immune regulation in the tumor microenvironment and its relevance in cancer therapy. Cellular & Molecular Immunology **19**, 1–2 (2022)