Supplementary File: A SURVIVAL ANALYSIS AND GRADE PREDICTION MODEL FOR LUNG SQUAMOUS CELL CARCINOMA BASED ON MULTIPLE INSTANCE LEARNING AND MULTI-SCALE TRANSFORMER

1. Introduction

Although the current MIL-based LUSC analysis methods have achieved significant success in the tasks for cancer grade prediction and survival analysis, there are still three limitations corresponding to the three phases aforementioned.[18,19] First, large-scale irrelevant instance inputs introduce substantial redundancy into the model, hindering the model from focusing on crucial tumor regions with highly discriminative representations. Megapixel-level WSI contains terrific empty or artifactual regions and the presence of epithelial, connective and adipose tissue unrelated to the tumor. The input of non-tumor information introduces significant redundancy into the prediction model and detracts from its ability to focus on highly discriminatory cancer tissues.[20-22] Second, the instance-level representation of a single scale can only capture image information at a specific resolution and lacks robustness in dealing with tumors of different sizes, shapes and types. Current MIL-based WSI analysis methods generally focus on single-scale image information at a specific magnification, while it is difficult to consider both local details and global context.[23-25] This may lead to models that fail to fully understand pathological features and their distribution. Meanwhile, feature extractors based on pre-training of natural images are unable to comprehensively obtain medically relevant contextual features. Third, simple instance-level feature aggregation strategy (e.g., concatenation) may raise dimensionality disaster and ignore the correlation and importance between instances. The aggregation strategy of large-scale instance-level features is critical to the predictive performance of the model, particularly the fusion of multi-scale features. In addition to concatenation, current methods typically employ the strategies of average or maximum pooling to fuse multi-scale features.[26,27] However, the pooling operation may lose certain detail information and also introduce redundant features, affecting the effectiveness of the model. Therefore, it is a challenge to effectively aggregate instance-level features to reduce feature spatial complexity while ensuring adequate fusion of multi-scale information.

In summary, the contributions of this paper are as follows:

- We investigate the application of the self-attention mechanism to multi-instance feature sampling for WSI, which improves the discriminability of the instance-level representation and reduces the redundancy of the model.
- We construct a multi-scale feature extractor based on self-supervised learning, which performs Convolution-Transformer based contextual feature extraction on WSI. The pathology representations at three scales comprehensively reflect the properties of the tumor microenvironment, structure of the tumor cells and polymorphism of the nuclei.
- In order to integrate multi-scale features more rationally, we develop a new multi-scale pathology feature aggregation strategy based on GRU. The utilization of sparse Transformer further reduces the redundancy of the features across scales.
- Extensive ablation experiments prove the validity of the proposed modules. Comparative results on the TCGA dataset LUSC and on multiple cross-cancer datasets demonstrate the validity and accuracy of the method. Interpretability studies also provide a basis for predicting results.

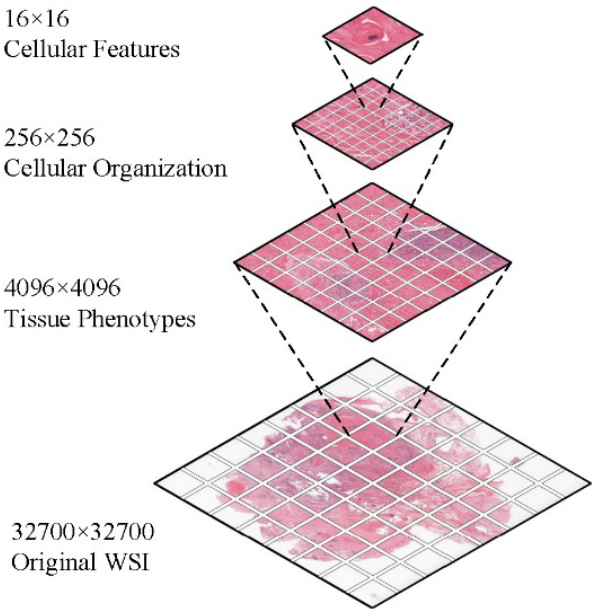


Fig. S1. Hierarchical Structure of Whole Slide Image.

2. Related work

2.1 Instance sampling strategy

Recently, some works utilize the attention mechanism to sample patches.[33,34] For example, BenTaieb et al.[33] utilize attention model to select patches for LUSC, but the training of the location model relies on manual annotations. Inspired by this, we propose a patch feature sampling strategy based on the self-attention mechanism in Transformer[35] to automatically select pathology representations with high discriminability without instance-level annotations.

2.2 Instance features extraction

Therefore, we propose a multi-scale pathology feature extractor based on self-supervised learning. Morphological feature representations of patches are efficiently extracted without additional annotations. The introduction of multi-scale learning provides a more detailed pathological analysis from cancer tissue regions to tumor cell features, which facilitates fine-grained prediction of cancer and improves the effectiveness and interpretability of the model.

2.3 Aggregation of instance features

However, HIPT ignores spatial redundancy between features at different scales, which may lead to an increase in the complexity of the model, thus impacting the predictive performance of downstream medical tasks. Therefore, we construct a multi-scale instance-level feature aggregation module based on gating units. The proposed instance aggregation strategy significantly improves the feature expression ability and generalization performance of the model by adaptively adjusting the weights of features at different scales. We also utilize the sparse Transformer44 to further reduce the information redundancy and improve the robustness of the model.

3. Methods

3.1 Problem formulation

The MIL typically crops patches from the WSI to obtain instances.[45] For a full resolution pathology image W_i, the localized block instances obtained from it are denoted as X_i, where $i \in (1, \dots, N)$ is the patient sequence number and j denotes the number of instances. In our work, WSI is divided into three scales of instance patches as X_{4096}^{ij}, X_{256}^{ij} and X_{16}^{ij}. Then the corresponding multi-scale feature

representations are obtained for each scale patches by the feature extractor, which is formulated as $F_{4096}^{ij} = E_l(X_{4096}^{ij})$, $F_{256}^{ij} = E_m(X_{256}^{ij})$ and $F_{16}^{ij} = E_s(X_{16}^{ij})$. $E()$ denotes the feature extractor at different scales. Bag is a collection of multi-scale instance features extracted from the same WSI, formulated as:

$$B_i = Aggre\left(F_{4096}^{ij}|_{j=1}^{L}, F_{256}^{ij}|_{j=1}^{M}, F_{16}^{ij}|_{j=1}^{S}\right) \quad (1)$$

where $Aggre()$ denotes a specific multi-scale instance features aggregation method. L, M and S indicate the total number of instances at different scales (large, medium and small). Further, each bag has a corresponding label, while the instances have no individual labels. The label of B_i is expressed as y_i, which is based on the overall features of the instances within the bag. For the grading prediction task for cancer diagnosis, $y_i \in \mathbb{R}^{1\times c}$, where c denotes the cancer grading stage. For the survival analysis of cancer prognosis, $y_i \in \mathbb{R}^{1\times 1}$ indicates the survival time of the patient. Based on the set of bag-level feature representations and labels, MIL next learns a downstream predictor for cancer diagnosis and prognosis.

3.1 Multi-scale feature extraction and sampling

Fig. S1 shows the visualization of pathology instances at different scales. With the proposed refined multi-scale segmentation, the model is able to sufficiently learn the hierarchical information of tissues, cell clusters and cells to improve the prediction performance of downstream tasks. For ease of representation, the number of instances j will be abbreviated in the following to denote this scale feature at the instance-level.

Evaluating the importance of input features via Transformer's multi-head self-attention mechanism has been widely used in natural image tasks,[49,50] demonstrating the effectiveness of this strategy for feature selection. However, the strategy has rarely been studied in diagnostic and prognostic in WSI. The proposed FSM is embedded into the integral training of the model in a plug-in style to select instances that are of interest for the downstream medical prediction task, thus reducing redundancy in the input space.

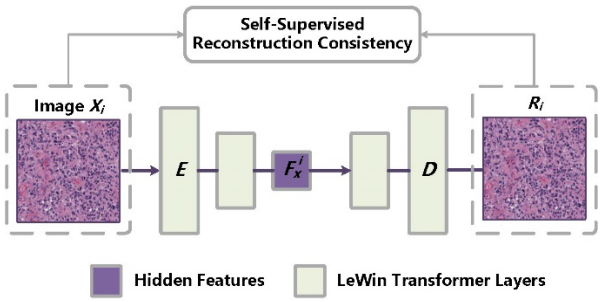


Fig. S2. Architecture of the proposed feature extraction module based on self-supervised learning.

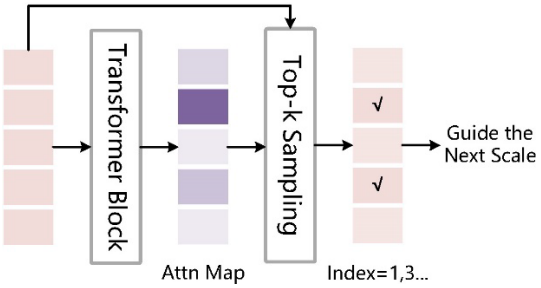


Fig. S3. Schematic of instance feature sampling based on self-attention mechanism.

3.2 Sparse encoder and multi-scale aggregation

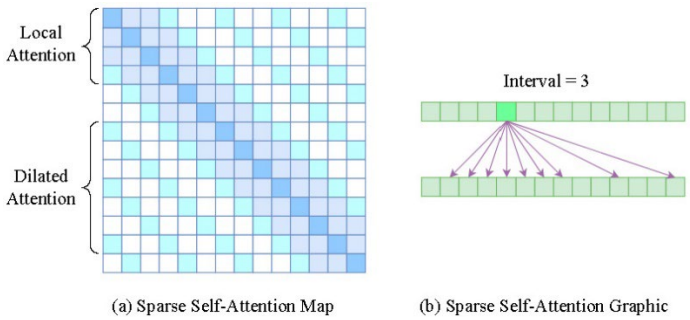


Fig. S4. Schematic of the sparse self-attention mechanism. (a) Combining local and dilated sparse self-attention map reduces the computational complexity. (b) Schematic diagram of sparse self-attention with interval 3.

Pre-extracted and sampled multi-scale features contain morphological information for each instance. However, different instances reflect diverse pathologic information of the corresponding cancer. For example, tumor-infiltrated regions typically exhibit a significant increase in cell density and a variety of cell morphologies without focusing on the color of the instance.[51] In addition, small-scale instances retain a great deal of feature information even after sampling.

Traditional multi-instance aggregation methods such as concatenation and pooling are unable to leverage the correlation between multi-scale features and ignore the variability of different features. Therefore, we propose MFAM to aggregate multi-scale features based on GRU. Through update gate and reset gate, MFAM dynamically models multi-scale hidden features for correlation. And the final aggregated output is constructed using the hidden feature that is relevant to the downstream task. The large and medium-scale instances are aggregated with the output $F^i_{4096-256} \in \mathbb{R}^{1\times k\times 256}$. We further fuse $F^i_{4096-256}$ with the small-scale instance features $F^i_{16-k'}$ to obtain the final multi-scale aggregated features $F^i_{agg} \in \mathbb{R}^{1\times k\times 256}$. In contrast to concatenation and pooling, MFAM introduces GRU-based nonlinear transformations to capture the deep relationships between multi-scale features and is able to dynamically select and emphasize important features, thus improving the effectiveness of instances aggregation.

3.3 Prediction

Grade prediction helps to improve treatment outcomes and patient survival by assessing the degree of malignancy and aggressiveness of a tumor and predicting a patient's risk of disease recurrence. For survival analysis of the aggregated features, the model utilizes full connection layers to obtain the final output $O^i_s = FC_s(F^i_{agg}) \in \mathbb{R}^{1\times 1}$. Suppose a patient is labeled as (t_i, σ_i), where t_i is survival time of and σ_i denotes whether the patient's survival time is censored or not. The negative log partial likelihood loss is:

$$L_s = -\sum_{i=1}^{N} \sigma_i \left(O^i_s - log \sum_{j:t_j \geq t_i} \exp\left(O^j_s\right) \right) \tag{7}$$

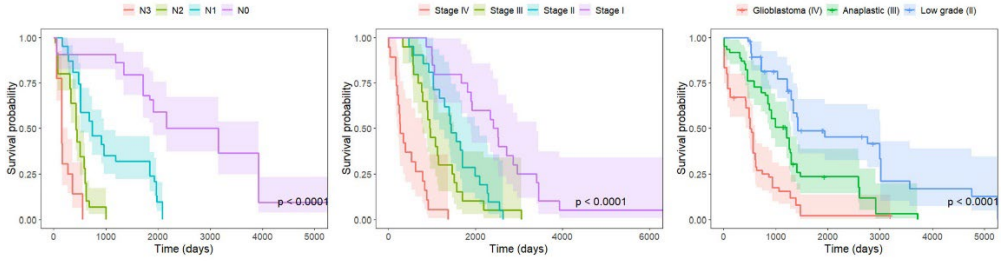
where j is the set of survival times equal to or larger than t_i. The negative log partial likelihood loss penalizes the risk prediction of the model and encourages the model to predict patient survival more accurately. For the cancer grade prediction task, we use cross-entropy loss as the optimization objective of the model, which is defined as:

$$L_g = -\sum_{i=1}^{N} \sum_{c=1}^{C} y^{ic}_g \log O^{ic}_g$$

where C and c denote the number of classes and the predictive class. $O^i_g = FC_g(F^i_{agg}) \in \mathbb{R}^{1\times C}$

represents the prediction output. y_g^{ic} is the one-hot label of the sample, which takes 1 if the true category of sample i is equal to c and 0 otherwise. O_g^{ic} denotes the predicted probability that sample i belongs to category c.

4. Experiments



(a) TCGA-LUSC (b) TCGA-BRCA (c) TCGA-GBMLGG

Fig. S5. The KM-estimation curves of our proposed MSTrans-MIL. The results of survival risk prediction across cancer types demonstrate the validity and scalability of the proposed method.

Table S2. Comparison results with SOTA methods in cancer grade prediction. Where 'AUC' is the area under the ROC curve, 'F1' denotes the reconciled average of precision and recall, 'Acc' is classification accuracy.

Model	LUSC			LUSC			LUSC		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
WSISA	0.915	0.842	0.874	0.852	0.722	0.795	0.881	0.805	0.813
CSCSP	0.926	0.869	0.896	0.875	0.773	0.854	0.905	0.819	0.836
DTMIL	0.943	0.882	0.917	0.883	0.784	0.861	0.919	0.832	0.847
TransMIL	0.944	0.885	0.913	0.867	0.776	0.854	0.895	0.812	0.824
DSMIL	0.941	0.879	0.911	0.889	0.782	0.865	0.927	0.844	0.852
HIPT	0.965	0.903	0.938	0.931	0.818	0.895	0.954	0.887	0.905
ZoomMIL	0.958	0.895	0.924	0.902	0.796	0.883	0.933	0.852	0.861
MG-Trans	0.976	0.912	0.945	0.938	0.820	0.906	**0.963**	0.896	0.915
MSTrans-MIL	**0.985**	**0.914**	**0.953**	**0.945**	**0.826**	**0.912**	0.961	**0.905**	**0.928**

4.1 Dataset and Implementation Details

We collected multiple datasets from the large-scale cancer program The Cancer Genome Atlas (TCGA) [53], a public cancer data consortium containing diagnostic WSI and corresponding survival outcomes and histological grade labels, to validate the effectiveness of the proposed method. The statistics of the data collected are shown in Tab. \ref{data}. In this study, we primarily investigate the survival analysis and grading prediction of lung squamous cell carcinoma (LUSC). In order to validate the scalability and generalization of the proposed model, we also investigate breast cancer invasive carcinoma (BRCA) and glioblastoma and low-grade glioma (GBMLGG) simultaneously. For each patient sample, we collected all diagnostic pathology images and obtained 1563 WSIs. For survival analysis and grading prediction, we used C-Index[54] and multiple classification metrics[55] to evaluate model performance, respectively.

4.2 Comparison results

In the single-scale MIL methods, WSISA employs a clustering-based patch sampling strategy that utilizes downstream task labels for supervised feature extraction for each class. This method is limited by the simple convolutional feature extractor and labeling requirements, which has weak model prediction performance and scalability. DTMIL and TransMIL further perform feature extraction of

instances based on a ResNet network pre-trained on natural images, but the semantic differences between different types of images are ignored. CSCSP introduces self-supervised learning with image coloring as a pretext task to extract instance features, which improves the effectiveness of instance feature extraction. However, it is limited by the inadequacy of single-scale instance features in representing pathology diagnostic information and unable to obtain satisfactory results. Notably, the introduction of the Transformer structure facilitates the modeling of WSI's pathology information at the global level, thus improving the predictive performance of the model.

Table S1. Data statistics. The datasets are collected from TCGA and 'M' denotes million. This contains the dataset division and the number of patches at each scale.

Dataset	LUSC	BRCA	GBMLGG
Number of WSI	326	255	982

4.3 Ablation study

Effects of Each Module in MSTrans-MIL. The specific module ablation settings are described as follows:

- Baseline (BL): Feature extraction is performed on typical 256×256 patch instances using ResNet-50 pre-trained on ImageNet. Only empty regions on the WSI are eliminated and all instance features are retained.
- BL + CT: Extraction of 256×256 instance features utilizing Convolution-Transformer based self-supervised learning strategy.
- BL + CT + MS: The multi-scale learning method is introduced, WSI is subdivided into 4096 ×4096 tissue regions, 256×256 tumor cellularity and 16×16 cellular features. Multi-scale instance features are aggregated and obtained by concatenation to F_agg^i∈R^(1×k×416).
- BL + CT + MS + FSM: The feature sampling module based on self-attention mechanism is introduced into the above method, where the number of Top-k features selected is set to 64.
- BL + CT + MS + FSM + MFAM: On the basis of the above, GRU-based feature aggregation strategy is utilized to dynamically fuse multi-scale instances.

However, the feature sampling module effectively reduces the input redundancy of the model, thus improving the computational efficiency. The use of sparse self-attention in the multi-scale feature aggregation module also effectively reduces the spatial redundancy of the model , which improves the prediction performance while reducing the computational complexity.

Effect of Scale Combination. However, the combination of the three scales further improves the model's learning of comprehensive biological implicit information from pathology images and obtains the best prediction performance. In addition, the fusion of pathology information at three scales has no significant increase in terms of computation and inference time, thus the proposed multi-scale feature fusion strategy is highly scalable.

4.4 Multi-scale interpretability

MSTrans-MIL is able to accurately identify regions of tumor tissue that exhibit good spatial concordance

with annotated tumor regions identified by professional pathologists. This shows that the proposed FSM is able to extract highly discriminative cancer regions well and effectively reduce the input redundancy that non-tumor regions impose on the model.

For the glioblastoma samples, medium-scale instances of tumor cellularity present large regions of necrosis with high tumor cell densities and significant nuclear heterogeneity, which are typical glioblastoma features.[62] The critical small-scale cellular morphologic features indicate multiple morphologically irregular nuclei, reflecting a high degree of cellular heterogeneity. Significant microvascular proliferation is also observed, indicating neovascularization of the tumor during invasion.[63] Through multi-scale instance interpretability analysis, we are able to discern the pathological properties of patients at different levels, thus facilitating the development of rational treatment plans, which also brings potential for AI-based precision medicine.