Spatial Encoding Techniques for Vision Transformer-Based Breast Histopathology Classification

İrem Damla Karagöz
Advisor: Prof. Selim Aksoy
Department of Computer Engineering
Bilkent University, Ankara, Türkiye

***Abstract*—Effective analysis of Whole Slide Images (WSIs) in breast histopathology requires models that can capture both local structures and broader spatial relationships. We investigate spatial encoding strategies for Vision Transformers using the BRACS dataset, where precise classification remains clinically challenging, especially of atypical cases. Leveraging Prov-GigaPath embeddings, we evaluated several tile ordering methods, including raster, spiral, and Hilbert curves, along with both absolute and relative positional encodings. Our results show that combining Hilbert traversal with 1D sinusoidal encoding, and spiral traversal with relative positional encoding, significantly enhances classification performance. These findings highlight the importance of spatial priors in improving transformer-based WSI interpretation.**

***Index Terms*—Vision Transformer, Whole Slide Image (WSI), Breast Histopathology, Positional Encoding, BRACS Dataset**

I. Introduction

Computational pathology involves the analysis of medical images, particularly Whole Slide Images (WSIs), high-resolution digital scans of tissue slides, to help diagnose diseases such as cancer, identify abnormalities, and predict disease outcomes. Its potential lies in improving cancer diagnostics by supporting various clinical tasks such as cancer subtyping, staging, and diagnostic/prognostic predictions [1].

Whole Slide Images (WSIs) present a unique challenge due to their large number of tiles and the need for context-aware interpretation of spatially distributed features. Proper analysis requires capturing both the local and global context to ensure that spatial dependencies and structural relationships between tiles are taken into account [2]. Transformer-based architectures have emerged as a powerful alternative to traditional convolutional neural networks (CNNs) for image analysis due to their capacity to capture long-range dependencies between image patches and identify important regions using their self-attention mechanism. This self-attention allows the model to weigh the relevance of different patches in relation to one another, making it highly effective for tasks requiring global context, such as classification [2], [3]. However, the effectiveness of self-attention in transformer-based models also heavily relies on how positional information is incorporated. To address this, positional information is embedded into the model to preserve the spatial arrangement of the patches and guide the attention mechanism accordingly [2].

In this study, we aim to investigate the impact of different patch ordering and positional encoding strategies on the performance of transformer-based whole slide image (WSI) classification using the BRACS dataset, with the goal of improving diagnostic accuracy in distinguishing between different breast cancer subtypes [4]. We use the Prov-GigaPath WSI tile encoder, which is specifically designed to process large-scale pathology [5]. We implemented and compared five distinct tile ordering strategies, each aiming to preserve or enhance spatial context in a different manner. This research is motivated by the critical need to improve diagnostic accuracy, particularly in identifying atypical cases that fall between benign and malignant classifications.

II. Related Work

A. Overcoming CNN Constraints with Transformer Models

Deep learning revolutionized medical image segmentation by enabling models to learn complex semantic features from data, improving both accuracy and adaptability across various medical imaging tasks [6]. Early successes were largely driven by Convolutional Neural Networks (CNNs), such as U-Net (ISBI 2015) and SegNet (CamVid), which use convolution and pooling to capture local spatial patterns [7], [8]. However, CNNs have trouble modeling global relationships and long-range dependencies within an image due to their local receptive fields, which limits performance on high-resolution images like Whole Slide Images (WSIs), where key features may be spread across large areas [9].

In the context of Whole Slide Image (WSI) analysis, the importance of capturing spatial context has been recognized. Campanella et al. (2019) demonstrated the effectiveness of attention-based deep learning for WSI classification, highlighting the model's ability to focus on diagnostically relevant regions. Their work, while not directly addressing positional encoding, underscores the necessity of spatially aware models in this domain [10]. Similarly, Li et al. (2021) introduced a context-aware attention network, which implicitly models spatial relationships to enhance WSI classification performance [1].

To address this, Transformers, introduced by Vaswani et al. (2017), have self-attention mechanisms that allow every part of the image to consider information from all other parts in a single step. The Vision Transformer (ViT) adapts the Transformer encoder for image classification. In ViT, the input image is divided into non-overlapping patches of a fixed size. Each patch is flattened into a one-dimensional vector along the channel dimension and mapped to a corresponding token through a linear projection. A class token is added to the image tokens, to aggregate the global features of the image and enable classification. Position embeddings are enriched with spatial information, ensuring the model understands the arrangement of patches. Embeddings are then fed into stacked Transformer encoders, where self-attention enables the model to capture long-range dependencies effectively [2].

B. Positional Encoding Strategies in Transformer Architectures

A widely adopted technique for conveying positional information to models is the use of positional encodings, which are combined with input embeddings. These encodings can be fixed functions of position (e.g., sinusoidal [2], [11]) or learned during training. Even CNNs, which capture local positional information, have been shown to benefit from positional encodings [12]]. Information about the relative or absolute position of tokens is given to the transformer architecture by adding positional encodings to the input embeddings at the base of both the encoder and decoder [2].

While absolute positional encodings help inform the model about position, they may fall short in capturing relative spatial relationships that are especially important in WSIs. To address this, Relation-aware Self-Attention was proposed by modifying the self-attention mechanism to incorporate relative distance between tokens directly [14]. Instead of relying solely on positional encodings, their method introduces trainable embeddings that represent the relative distance between positions in the input sequence. These embeddings are added to the key and value vectors in the self-attention mechanism, allowing the model to learn spatial relationships such as "how far" and "in what direction" one tile is from another.

Additionally, this approach of applying systematic tile ordering methods aligns with observed visual search behaviors in histopathology, where it is known that pathologists often employ systematic scanning strategies [15].

III. Methodology

This section details the experimental setup and procedures employed in this study. An overview of the pipeline is provided, followed by descriptions of the tile ordering methods, the Vision Transformer classifier architecture, and the loss function and optimization strategy. Finally, the dataset and the pre-trained feature extraction model are described.

A. Project Pipeline

1) *Tile extraction:* Whole-Slide Images (WSIs) from the BRACS dataset were processed to extract tiles of size 256×256 pixels at $10\times$ magnification [4].
2) *Embedding generation:* The extracted tiles were passed through the Prov-GigaPath pre-trained model to generate 1536-dimensional feature vectors (embeddings) for each tile [5].
3) *Tile ordering:* The resulting tile embeddings for each WSI were reordered using five different strategies to incorporate spatial information.
4) *Classification:* The ordered sequences of tile embeddings were fed into a Vision Transformer (ViT) classifier to predict the lesion type.
5) *Performance evaluation:* Classification performance was assessed using standard metrics including accuracy, precision, recall, and F1-score.

B. Dataset Preparation

We utilized the BRACS dataset for all experiments, which provides Whole Slide Images (WSIs) annotated for various breast lesion types. In detail, the dataset contains 547 Whole-Slide Images (WSIs) and 4539 Regions of Interest (ROIs), each carefully annotated by board-certified pathologists to ensure accurate labeling. The annotations divide the lesions into three types: benign, malignant, and atypical, which are further subdivided into seven specific subtypes [4]. In this project, we have not yet focused on the categorization of these subtypes, as we aim to increase the accuracy of benign/malignant/atypical classification before delving into finer-grained distinctions, and due to the challenges of subtype classification.

The dataset has high class imbalance across the three categories. For example, in the training set, there are 202 benign, 140 malignant, and only 52 atypical samples. Similar distributions are observed in the validation (30/21/14) and test (32/32/22) splits. To mitigate this imbalance, we apply class weighting in the loss function to ensure equitable learning across categories.

To convert the WSIs into a format suitable for transformer-based learning, we applied tile-based preprocessing and embedding extraction. Before tiling, we applied a luminance threshold to isolate tissue regions and eliminate background areas lacking diagnostically relevant content in each WSI. Later the images were divided into non-overlapping tiles of 256×256 pixels at 10x magnification (level 2) (See Fig. 1 for an example WSI before and after the tiling process.). This magnification level was preferred to balance computational efficiency with sufficient histological detail. We used Prov-GigaPath, a pre-trained foundation model specifically designed for digital pathology, to extract meaningful feature embeddings from individual tiles of large medical images (WSIs). To optimize processing and focus on relevant areas, Prov-GigaPath automatically discards tiles with less than 10% tissue coverage to eliminate irrelevant background regions and reduce computational overhead. The Gigapath tile encoder produces a 1536-dimensional embedding vector, representing the features of each tile [5].

All embedding generation was performed on Google Colab using NVIDIA Tesla T4 GPUs to accelerate computation. The result was a dataset of numerical embedding sequences

corresponding to each WSI, serving as input to the Vision Transformer models tested in this study.

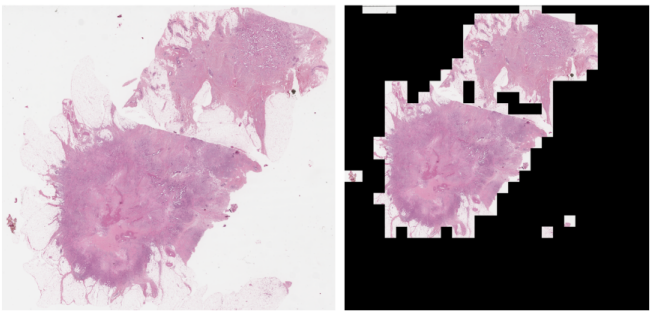

Fig. 1. Visualization of a whole slide image (WSI) before and after tile extraction. Non-overlapping tiles are extracted and recombined for illustration.

C. Tile Ordering and Positional Encoding Strategies

1) 1D Sinusodial Encoding: We use 1D sinusoidal positional encodings which was introduced by Vaswani et al. [2]. The positional encoding vector for position pos and dimension i is defined as:

$$\text{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{1}$$

$$\text{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2}$$

where d is the embedding dimension.

In our implementation, given a sequence length L and embedding dimension d (embedding vector size), we generate a matrix of shape $L \times d$, where each row corresponds to the positional embedding of one tile. These vectors are then added to the tile embeddings before being fed to the transformer encoder.

2) 2D Sinusodial Encoding: This technique is an extension of 1D positional encoding initially introduced by Vaswani et al. [2]. In 2D sinusoidal positional encoding, we compute each coordinate axis separately, each into embedding dimension $d//2$ dimensions, then concatenate to form the final embedding of size embedding dimension d. instead of creating the embedding dimension d is split between the x (horizontal) and y (vertical) coordinates. Each coordinate is encoded using the 1D sinusoidal formulas. We then concatenate the two encodings to obtain a single vector, resulting in a position embedding matrix of shape $[H \times W, D]$, where H and W are the grid height and width, and D is the embedding dimension. This approach, widely adopted in vision transformers such as ViT [16], enables the model to distinguish between different spatial locations

3) Relative Positional Encoding: Our implementation of relative positional encoding was created using the formulation by Shaw et al. (2018) [14], and is also inspired by Hugging Face's T5 transformers library which similarly applies relational bias to text [17]. We defined a custom RelativeMultiHeadAttention module that adds learnable relative position biases to the attention scores. This is wrapped in a transformer encoder layer with RPE, which replaces the standard encoder layer in the WSITransformer.

We paired positional encoding strategies with various tile ordering methods to obtain the configurations below:

- *Raster Scan (`rasterscan`)*
  Arranges tiles in row-major order, traversing from left to right across each row and top to bottom across rows. Raster order is implemented by lexicographically sorting tile coordinates. It serves as a baseline ordering with minimal spatial awareness.

- *Raster Scan with 1D Positional Encoding (`rasterscanwencoding`)*
  Uses the same row-major ordering as `rasterscan`, but adds a 1D sinusoidal positional encoding to each tile embedding based on its sequence index.

- *Spiral Scan (`spiral`)*
  Reorganizes tiles to follow a spiral path inward from the outer margins of the tissue.

- *Spiral with 1D Positional Encoding (`spiralwencoding`)*
  Retains spiral ordering and adds 1D sinusoidal positional encodings to represent the sequence index. This provides explicit positional context to the transformer, reducing reliance on learned attention patterns alone.

- *Hilbert Curve with 1D Positional Encoding (`hilbertwencoding`)*
  This method orders tiles along a Hilbert curve ,a space-filling fractal that maximally preserves spatial locality. Hilbert curves mathematically optimize for spatial locality preservation in 1D traversals, minimizing the average distance between neighbors in 2D space when mapped into a linear sequence [18]. Preserving adjacency is expected to improve the transformer's ability to model neighborhood structures, gland formations, or infiltrative patterns, which are critical for histopathological classification. The addition of 1D positional encoding further reinforces positional awareness, potentially enabling the model to integrate both absolute and relative positional cues.

- *2D Sinusoidal Encoding (`2dsinusodial`)*
  Unlike prior methods that reorder tiles, this method retains original coordinate-based ordering but injects a 2D sinusoidal positional embedding directly encoding (x, y) coordinates into each tile embedding.

- *Relative Positional Encoding*

The RPE mechanism is applied independently of tile ordering strategies, hence we selected order methods that do not apply 1D positional encoding: `hilbert`, `rasterscan`) and `spiral`).

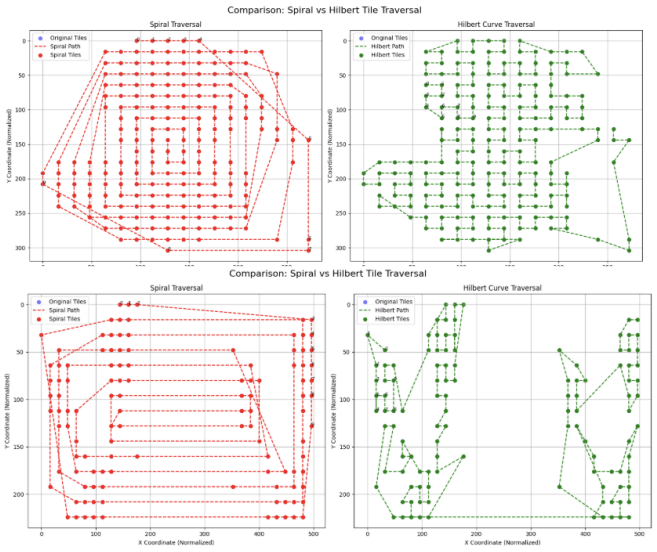

Fig. 2. Visualization of tile traversal techniques spiral and Hilbert curve.[1]

IV. Experimental Setup

A. Hardware and Software Environments

All experiments were performed on Google Colab using an NVIDIA Tesla T4 (L4) GPU. The environment included a Linux-based OS, Python 3.8, and PyTorch 1.12 for model development. Matplotlib was used for visualizing training and validation metrics. The cloud-based setup enabled efficient processing of high-dimensional WSI data without local computational resources.

B. ViT Architecture and Model Training

For this research, we implemented and used a custom transformer model designed for whole slide image (WSI) classification into three diagnostic categories: benign, atypical, and malignant. Each WSI is preprocessed into up to 600 tile embeddings (1536-dim), which are spatially reordered using techniques such as raster scan, spiral, or Hilbert curve, optionally augmented with *positional encodings*.

In configurations using positional encoding, either 1D or 2D sinusoidal encodings are added to the tile embeddings to provide spatial context. The contribution of these positional encodings is scaled by a tunable hyperparameter α, allowing us to modulate their influence on the model. Specifically, the positional encodings (normalized between -1 and 1) are multiplied by α before being added to the tile embeddings (typically ranging between -8 and 8). Lower values of α reduce the impact of positional information, while higher values amplify it.

Choosing an appropriate α is critical: if the positional encodings are scaled too aggressively, they may overwhelm the semantic content encoded in the tile embeddings. Conversely, if α is too small, spatial information may be underrepresented. Since the embeddings contain rich information about the visual and morphological features of each tile, it is essential to maintain a proper balance to avoid diluting their discriminative power.

PyTorch DataLoader objects are used to efficiently handle batching of BATCH_SIZE = 16 and shuffling. Our model features a CLS token, positional encoding, a Transformer encoder with a Feed-Forward Dimension of 2048, and an MLP classification head designed for 3 classes: benign, atypical, malignant. Focal loss and the Adam optimizer, with a learning rate of 1e-4 learning rate and learning rate scheduler, are used for training. The model trains for 70 epochs over the entire dataset, with validation performed after each epoch to monitor performance and prevent overfitting. The best-performing model (best validation accuracy) is saved. Finally, the trained model is evaluated on the test set, with performance metrics such as loss, accuracy, precision, recall, and F1-score calculated and printed. A confusion matrix is generated, and plots of training and validation metrics are created.

To address class imbalance during training, we apply class weighting within the focal loss function. Class weights are computed using `compute_class_weight` from scikit-learn, assigning higher weights to underrepresented classes based on their inverse frequency. These weights are passed as the α parameter to the focal loss, which extends cross-entropy by incorporating a modulating factor $(1 - p_t)^\gamma$ to emphasize hard-to-classify examples [19]. This weighted formulation encourages the model to focus on minority classes, which is particularly important in histopathological datasets where clinically significant categories (e.g., malignant) may be underrepresented.

V. Evaluation Metrics

A. Confusion Matrix Interpretation

In the confusion matrix, each element $cm[i][j]$ represents the number of samples whose true class is j and that were predicted as class i. We can interpret the elements as follows:

Diagonal elements represent correct classifications (true positives for each class). Off-diagonal elements represent misclassifications, and $cm[i][j]$ where $i \neq j$, represents the false positive (FP) count for class i and false negative (FN) count for class j.

The confusion matrix which summarizes prediction results can be expressed as:

$$\text{Confusion Matrix} = \begin{bmatrix} \text{TP}_0 & \text{FP}_0, \text{FN}_1 & \text{FP}_0, \text{FN}_2 \\ \text{FP}_1, \text{FN}_0 & \text{TP}_{11} & \text{FP}_1, \text{FN}_2 \\ \text{FP}_2, \text{FN}_0 & \text{FP}_2, \text{FN}_1 & \text{TP}_{22} \end{bmatrix} \tag{3}$$

[1]The plots were generated using a custom Python function with Matplotlib, based on traversal coordinates computed during embedding reordering.

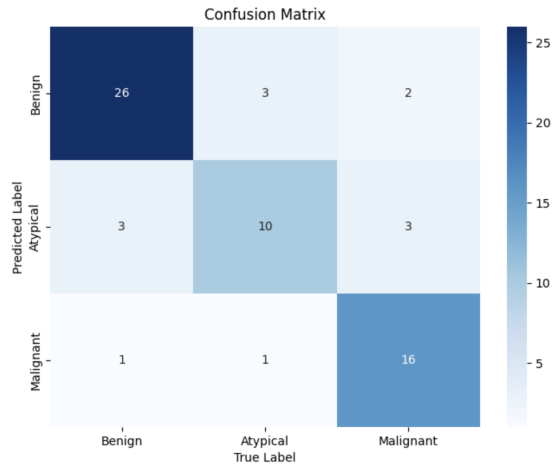

Fig. 3. Confusion matrix obtained for *Spiral* with RPE configuration (α=1).

B. Metrics

Detailed formulation of the metrics are presented below:

1) Accuracy: Accuracy is the ratio of correctly predicted samples to the total number of samples:

$$\text{Accuracy} = \frac{\text{Total correct predictions}}{\text{All predictions}} \tag{4}$$

Which refers to the confusion matrix elements as below, where C is the number of classes:

$$\text{Accuracy} = \frac{cm[0][0] + cm[1][1] + ... + cm[C-1][C-1]}{\sum_{i=0}^{C-1}\sum_{j=0}^{C-1} cm[i][j]} \tag{5}$$

2) Precision (Macro-Averaged): Precision measures the proportion of true positive predictions among all positive predictions. Macro-averaged precision is the average of precision for all classes.

$$\text{Prec.}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \quad \text{Macro Prec.} = \frac{1}{C}\sum_{c=1}^{C}\text{Prec.}_c \tag{6}$$

Which refers to the confusion matrix elements as below:

$$\text{Prec.}_c = \frac{cm[c][c]}{\sum_{j=0}^{C-1} cm[c][j]} \tag{7}$$

$$\text{Macro Prec.} = \frac{1}{C}\sum_{c=0}^{C-1}\frac{cm[c][c]}{\sum_{j=0}^{C-1} cm[c][j]} \tag{8}$$

3) Recall (Macro-Averaged): Recall measures the proportion of actual positives correctly predicted:

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \quad \text{Macro Recall} = \frac{1}{C}\sum_{c=1}^{C}\text{Recall}_c \tag{9}$$

Which refers to the confusion matrix elements as below:

$$\text{Recall}_c = \frac{cm[c][c]}{\sum i = 0^{C-1} cm[i][c]} \tag{10}$$

$$\text{Macro Recall} = \frac{1}{C}\sum_{c=0}^{C-1}\frac{cm[c][c]}{\sum_{i=0}^{C-1} cm[i][c]} \tag{11}$$

4) F1-Score (Macro-Averaged): The final F1-score metric is the average of the per-class F1-scores, which are harmonic mean of the macro-averaged precision and recall.

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \quad \text{Macro F1} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c \tag{12}$$

VI. Results and Discussion

A. Published Baseline

TABLE I
RESULTS AT WSI LEVEL FOR 3-CLASS CLASSIFICATION TASK FROM BRACS BASELINE [4].

	Benign	**Atypical**	**Malignant**	**Total**
F-measure	74.4	57.2	78.0	69.8
Precision	75.6	51.5	86.5	71.2
Recall	72.5	65.2	71.9	69.9
Accuracy	–	–	–	70.3

Table I presents class-wise evaluation metrics reported at the WSI level for the BRACS dataset as published by Brancati *et al.* [4]. These values serve as a performance baseline for comparison with our model.

B. Comparative Analysis of Results

TABLE II
AVERAGE PERFORMANCE METRICS BY ORDER AND ALPHA

Order	**Alpha**	**Val. Acc.**	**Test Loss**	**Test Acc.**	**Prec.**	**Recall**	**F1**
2dsinusodial	1	0.69	2.41	0.67	0.55	0.58	0.54
hilbertwencoding	0.2	0.70	2.12	0.70	0.63	0.62	0.60
hilbertwencoding	0.3	0.68	1.49	0.68	0.64	0.60	0.60
hilbertwencoding	0.4	0.69	1.45	**0.77**	0.76	0.74	**0.75**
hilbertwencoding	0.5	0.69	1.97	0.69	0.67	0.61	0.60
hilbertwencoding	1	0.68	1.87	0.66	0.60	0.57	0.56
rasterscan	1	0.70	1.00	0.70	0.72	0.65	0.66
spiral	1	0.71	1.57	0.72	0.67	0.64	0.65
rasterscanwencoding	0.2	0.71	2.00	0.75	0.73	0.69	0.70
rasterscanwencoding	0.3	0.72	2.65	0.71	0.62	0.62	0.60
rasterscanwencoding	0.4	0.70	1.44	0.68	0.60	0.61	0.60
rasterscanwencoding	0.5	0.69	0.85	0.76	0.76	0.68	0.67
rasterscanwencoding	1	0.68	1.30	0.73	0.72	0.68	0.69
spiralwencoding	0.2	0.72	1.61	0.70	0.67	0.63	0.63
spiralwencoding	0.3	0.73	1.40	0.74	0.74	0.69	0.70
spiralwencoding	0.4	0.73	1.19	0.72	0.67	0.66	0.66
spiralwencoding	0.5	0.74	1.07	0.69	0.67	0.66	0.66
spiralwencoding	1	0.69	1.03	0.68	0.59	0.60	0.58

We evaluated various tile ordering strategies, with and without absolute positional encoding, to evaluate their impact on classification performance. Table II summarizes the average validation accuracy, test metrics, and F1-scores across different order-encoding configurations. The encoding methods used in this table are merely absolute encoding methods.

The best performance was achieved by the *hilbertwencoding* method with $\alpha = 0.4$, which reached the highest test accuracy (0.77) and F1-score (0.75). This approach leverages the Hilbert

space-filling curve to preserve local spatial structure and also augments each tile embedding with 1D sinusoidal positional encoding. The results indicate that combining spatial location with explicit sequence information significantly enhances the ability of the model to capture histological patterns.

Hilbert curve ordering, which is optimized for preserving spatial locality via a space-filling fractal path, consistently outperformed other methods when combined with 1D positional encoding (hilbertwencoding). With $\alpha = 0.4$, this method achieved the highest test accuracy (0.77) and F1-score (0.75), highlighting the benefit of minimizing long-range jumps in the input sequence and preserving 2D adjacency information.

Spiralwencoding, also achieved strong performance across multiple configurations. Notably, with $\alpha = 0.3$, it yielded a test accuracy of 0.74 and an F1-score of 0.70. The consistent results across various α values suggest that spiral ordering provides a robust spatial prior that is further improved by the addition of positional encoding.

Raster scan with encoding ($\alpha = 1$) reached a test accuracy of 0.73 and F1-score of 0.69, outperforming the plain raster scan baseline (F1: 0.66). This demonstrates that even simple traversal patterns can benefit substantially from positional encoding, which compensates for their lack of spatial awareness.

In contrast, the *2dsinusodial* method, which directly injects 2D coordinate-based positional embeddings without reordering tiles, achieved the lowest test accuracy (0.67) and F1-score (0.54). We hypothesize that without an imposed traversal pattern, transformers struggle to exploit 2D spatial relations from positional embeddings alone.

Overall, our experiments supported the hypothesis that ordering methods, when combined with positional encoding, significantly improve transformer performance on WSIs. Among them, Hilbert curve tile traversal with positional encoding ($\alpha = 0.4$) yields the best generalization, while Spiral with positional encoding offers biologically aligned ordering with strong accuracy. The unordered tiles (rasterscan) performance was behind spiral and hilbert traversals, despite some gains from encoding. Many configurations resulted in higher performance than the baseline, with higher F1 scores than 69.8 and higher accuracy than 70.3 (see tables I and II) [4].

TABLE III
AVERAGE METRICS FOR SELECTED METHODS WITH RELATIVE POSITIONAL ENCODING

Order	Alpha	Val. Acc.	Test Loss	Test Acc.	Prec.	Recall	F1
hilbert	0.2	0.72	0.86	0.75	0.76	0.71	0.73
hilbert	0.3	0.70	1.07	0.75	0.72	0.70	0.70
hilbert	0.4	0.70	0.87	0.80	0.79	0.76	0.77
hilbert	0.5	0.71	0.95	0.75	0.71	0.69	0.69
hilbert	1.0	0.70	1.22	0.72	0.69	0.66	0.66
rasterscan	0.2	0.73	1.51	0.75	0.72	0.70	0.70
rasterscan	0.3	0.72	1.27	0.77	0.79	0.70	0.71
rasterscan	0.4	0.71	2.11	0.75	0.76	0.66	0.66
rasterscan	0.5	0.70	1.06	0.69	0.66	0.63	0.64
rasterscan	1.0	0.73	1.82	0.74	0.72	0.68	0.69
spiral	0.2	0.71	1.55	0.74	0.70	0.68	0.68
spiral	0.3	0.72	1.37	0.69	0.62	0.61	0.61
spiral	0.4	0.69	0.68	0.69	0.73	0.70	0.68
spiral	0.5	0.74	1.10	0.72	0.68	0.67	0.68
spiral	1.0	0.72	0.81	0.80	0.78	0.77	0.77

C. Relative Positional Encoding Results Analysis

We evaluated three tile ordering strategies, *hilbert*, *rasterscan*, and *spiral*, applying relative positional encoding (RPE). The scalar hyperparameter α controlling the influence of RPE and the results are summarized in Table III.

The *spiral* and *hilbert* methods achieved the highest test accuracy (0.80) and F1-score (0.77). This indicates that traversal patterns work well when enriched with relative positional cues. The *spiral* with $\alpha = 1$ resulted in superior recall (0.77) and F1 score (0.77) also suggests that the model is sensitive to harder-to-classify categories like atypical lesions. Similar high performance was obtained using $\alpha = 0.4$, achieving the highest test accuracy (0.80) and the F1 score (0.77). However, performance consistently declined when α deviated from 0.4 in either direction.This aligns with the absolute positional encoding results, where $\alpha = 0.4$ often yielded the best performance.

The *rasterscan* configuration, despite its simplicity (linear row-major order), benefited notably from RPE, achieving 0.74 test accuracy and an F1-score of 0.69. This demonstrates that relative encoding can enhance even naive spatial arrangements by enabling the model to exploit neighborhood context. A similar trend was observed with the alpha values in the *rasterscan* order, as moderate values around $\alpha = 0.3$ and 0.3 led to improved metrics, while higher α values generally resulted in reduced performance.

Spiral displayed a different pattern than *rasterscan* and *hilbert* , achieving its best F1-score (0.77) and accuracy (0.80) at $\alpha = 1.0$, indicating it may benefit from stronger positional knowledge. While this was an exception, in general moderate α values offer a good trade-off between spatial context and feature representation.

The *hilbert* configuration, which uses a space-filling fractal curve to preserve spatial locality, resulted in lower performance when $\alpha = 1$ (0.72 test accuracy, 0.66 F1). While this is still an improvement from *hilbertwencoding with* $\alpha = 4$, was behind *rasterscan* and *rasterscan*. This suggests that while Hilbert curves maximize spaciality, they may not provide a traversal path that aligns well with how transformers learn relational context. Nontheless, it was observed that the results of relative positional encoding mostly overrule the absolute positional results (compare Table II *methodXwencoding* with Table III *methodX* results).

Overall, the experiments confirm that relative positional encoding improves classification performance, but the effectiveness depends on the compatibility between the ordering method and the model's attention mechanism. Among all tested combinations, *spiral + RPE ($\alpha = 1$)* and *hilbert + RPE ($\alpha = 0.4$)* proved to be the most effective methods.

VII. CONCLUSION

This work presents an evaluation of spatial encoding techniques for Vision Transformer-based classification of breast histopathology Whole Slide Images (WSIs) using the BRACS dataset. By integrating various tile traversal strategies with absolute and relative positional encoding, we systematically

investigated their impact on model performance. Results demonstrate that spatially informed traversal patterns such as Hilbert and spiral, when combined with positional encodings significantly enhance transformer-based WSI classification performance. We observed that spiral ordering with relative encoding and hilbert traversal with $alpha$-scaled relative encoding yielded the highest F1-scores (0.77), indicating improved sensitivity to challenging lesion types like atypical cases. Many configurations resulted in higher performance than the baseline, with higher F1 scores than 69.8 and higher accuracy than 70.3, proving (see tables I, II and III) [4].

Our findings emphasize the importance of incorporating spatial priors into WSI analysis and recommend that future research investigate more refined ordering methods and positional encoding schemes to further enhance transformer-based classification.

VIII. Acknowledgement

The author thanks Prof. Selim Aksoy of İhsan Doğramacı Bilkent University for his supervision, mentorship, and insightful discussions during weekly meetings throughout the course of this project.

References

[1] X. Li, C. Li, M. M. Rahaman, *et al.*, "A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches," *Artificial Intelligence Review*, vol. 55, pp. 4809–4878, 2022, doi: 10.1007/s10462-021-10121-0.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] C. Erkan and S. Aksoy, "Space-filling curves for modeling spatial context in transformer-based whole slide image classification," in *Medical Imaging 2023: Digital and Computational Pathology*, vol. 12471, SPIE, 2023.

[4] N. Brancati, A. M. Anniciello, P. Pati, D. Riccio, G. Scognamiglio, G. Jaume, G. De Pietro, M. Di Bonito, A. Foncubierta, G. Botti, M. Gabrani, F. Feroce, and M. Frucci, "BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images," *Database*, vol. 2022, 2022, Art. no. baac093, doi: 10.1093/database/baac093.

[5] H. Xu, N. Usuyama, J. Bagga, *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, pp. 181–188, 2024, doi: 10.1038/s41586-024-07441-w.

[6] Md. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, 2024, Art. no. 101504, doi: 10.1016/j.imu.2024.101504.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th Int. Conf.*, Munich, Germany, Oct. 5–9, 2015, Proc., Part III, vol. 9351, pp. 234–241, Springer, 2015.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[9] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *arXiv preprint arXiv:1701.04128*, 2017.

[10] G. Campanella, M. G. Hanna, L. Geneslaw, *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, pp. 1301–1309, 2019, doi: 10.1038/s41591-019-0508-1.

[11] S. Sukhbaatar, J. Weston, R. Fergus, *et al.*, "End-To-End Memory Networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[12] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. ICML*, 2017.

[13] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proc. EMNLP*, 2016.

[14] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. NAACL-HLT*, pp. 464–468, 2018.

[15] M. A. Mello-Thoms, D. J. Grzybicki, and J. M. Wilbur, "Eye tracking in digital pathology: Literature review and opportunities," *Journal of Pathology Informatics*, vol. 15, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11168484/

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] Hugging Face, "transformers/models/t5/modeling_t5.py," GitHub, Available: https://github.com/huggingface/transformers/blob/main/src/transformers/models/t5/modeling_t5.py (Accessed May 16, 2025).

[18] D. Hilbert, "Ueber die stetige Abbildung einer Linie auf ein Flächenstück," *Mathematische Annalen*, vol. 38, no. 3, pp. 459–460, 1891. doi: 10.1007/BF01199431.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.