Innovation Series https://www.innoseries.com

# GSDP-ViT: A Lightweight Vision Transformer for Cervical Histopathological Image Classification

**Liping Peng**
College of Computer and Information Science, Chongqing Normal University, Chongqing, China

**Abstract:** Cervical cancer is the fourth most common cancer among women worldwide, and histopathological image analysis remains the gold standard for diagnosing cervical precancerous lesions. However, the high morphological similarity among different lesion subtypes and the significant information sparsity in pathological microscopic images bring considerable challenges for automated classification. Current Vision Transformer (ViT) models applied to this task are limited by redundant feature generation in the tokenization stage, the lack of frequency-domain texture feature extraction, and wasteful computational allocation to uninformative image regions. To overcome these limitations, this paper proposes GSDP-ViT, a lightweight Vision Transformer that incorporates three novel components. First, a Ghost Convolutional Tokenizer (GCT) generates diverse feature maps from fewer parameters by augmenting intrinsic features through cheap depth wise linear operations. Second, a Spectral–Spatial Dual Attention (SSDA) mechanism simultaneously captures spatial morphological patterns and frequency-domain textural features via Discrete Cosine Transform (DCT) based parallel attention, with a learnable gating mechanism for adaptive fusion. Third, a Progressive Token Pruning (PTP) strategy dynamically evaluates token importance at each Transformer layer and removes the least informative tokens layer by layer, concentrating computation on pathologically relevant regions. Experiments on the LDCH cervical histopathological image dataset show that GSDP-ViT achieves 87.85% accuracy and 68.21% macro-F1 with only 0.47M parameters and 0.89 GFLOPs, surpassing several state-of-the-art models while maintaining superior computational efficiency. Ablation experiments validate the effectiveness and synergistic benefits of each component.

**Keywords:** Cervical histopathological images; Ghost convolution; Spectral–spatial dual attention; Dynamic token pruning; Lightweight Vision Transformer

## 1. Introduction

Cervical lesions constitute a spectrum of diseases that progress from mild atypical hyperplasia to invasive cervical cancer [1]. Among cervical precancerous lesions, cervical intraepithelial neoplasia (CIN) is the most prevalent and well-characterized type, accounting for roughly 95% of all cervical epithelial precancerous lesions [2]. These precancerous lesions can be detected and treated through established clinical procedures, thereby substantially lowering the risk of progression to invasive cancer [3]. In current clinical practice, pathologists perform histopathological examination to determine lesion grade by observing cell and tissue morphology under optical microscopy, in conjunction with clinical information such as patient age, HPV testing status, and sampling site [4]. This examination is widely considered the gold standard for cervical lesion diagnosis. Therefore, improving the accuracy and efficiency of cervical precancerous lesion classification carries critical importance for patient outcomes and clinical workflow.

Deep learning methods for image classification can be broadly categorized into convolutional neural network (CNN) based approaches and Vision Transformer (ViT) based approaches [5,6]. CNNs produce discriminative low-dimensional feature representations through hierarchical convolution operations, effectively capturing local structural patterns that are often overlooked during routine manual examination [7]. Vision Transformers, on the other hand, possess an inherent capability to model long-range dependencies across the full receptive field of medical images through their self-attention mechanism [6]. To combine these complementary strengths, researchers have proposed various hybrid models that integrate CNN and ViT components [8,9]. Although these methods have shown effectiveness in general image classification tasks, they still encounter three specific difficulties when applied to cervical histopathological image classification.

The first difficulty lies in redundant feature generation during the tokenization process. Morphological features among different pathological types in cervical histopathological images are often highly similar in localized regions. Standard convolutional tokenizers apply uniform convolution operations across all input channels, producing a large proportion of redundant feature maps that contribute little to distinguishing pathological subtypes. On relatively small medical datasets such as LDCH, this parameter redundancy not only wastes computational resources but also raises the risk of overfitting, ultimately constraining classification accuracy.

The second difficulty concerns the inability to simultaneously exploit spatial and frequency-domain information. Key diagnostic features in cervical histopathology, including chromatin texture coarseness, nuclear membrane regularity, and mitotic figure periodicity, exhibit distinctive characteristics in both the spatial domain and the frequency domain. Standard self-attention mechanisms in ViT operate solely on spatial token embeddings and fail to capture frequency-specific textural patterns that experienced pathologists implicitly assess when grading lesions. This limitation becomes particularly problematic when distinguishing morphologically similar but pathologically different subtypes such as CIN II and CIN III.

The third difficulty relates to wasteful computation on uninformative regions. Cervical histopathological images typically contain large areas of background, normal epithelium, or stroma that carry limited diagnostic value. Standard ViT models process all image tokens equally throughout the network, spending significant computation on irrelevant regions rather than concentrating capacity on pathologically meaningful areas. This uniform processing not only increases computational cost but also introduces noise that may degrade classification performance.

To address these three difficulties, this paper proposes GSDP-ViT, a lightweight single-branch Vision Transformer designed for cervical histopathological image classification. The main contributions are summarized as follows:

(1) A Ghost Convolutional Tokenizer (GCT) is proposed to replace standard convolutional patch embedding. The GCT first generates a compact set of intrinsic feature maps via standard convolution, then augments them through cheap depthwise linear transformations. This design reduces the tokenizer parameter count by roughly 50% while maintaining feature diversity.

(2) A Spectral–Spatial Dual Attention (SSDA) mechanism is designed to extend traditional self-attention with a parallel frequency-domain attention path. The spatial path captures structural

morphological patterns, while the spectral path applies 1D DCT to transform tokens into the frequency domain and performs lightweight linear attention on frequency components. A learnable gating mechanism adaptively fuses the two paths, enabling the model to leverage both spatial layout and textural frequency patterns.

(3) A Progressive Token Pruning (PTP) strategy is introduced to dynamically evaluate token importance after every Transformer layer and progressively prune the least informative tokens. This mechanism reduces computational cost in deeper layers while ensuring that the model focuses its representational capacity on pathologically relevant regions.

Experimental results on the LDCH dataset confirm that GSDP-ViT achieves state-of-the-art classification performance with only 0.47 million parameters, validating its effectiveness and efficiency.

## 2. Related Work

### *2.1 Vision Transformers for Medical Image Classification*

Since Dosovitskiy et al. introduced the Vision Transformer for image recognition [6], numerous variants have been proposed to improve efficiency and applicability in medical imaging. The Swin Transformer employs shifted window based self-attention to achieve linear computational complexity while maintaining hierarchical feature extraction [13]. DeiT introduces knowledge distillation strategies for data-efficient ViT training on smaller datasets [8]. T2T-ViT progressively tokenizes images through token-to-token transformations, capturing fine-grained structural information more effectively than flat tokenization approaches [19]. CMT and CSWin Transformer further incorporate convolutional operations and cross-shaped window attention to strengthen local feature extraction within the Transformer framework [15,16].

In the medical imaging domain, several specialized architectures have been developed. The Compact Convolutional Transformer (CCT) replaces the linear projection of original ViT with convolutional tokenization and introduces sequence pooling, achieving competitive performance with considerably fewer parameters [23]. The Enhanced Vision Transformer (EVT) builds upon CCT by incorporating wavelet positional embedding to reduce aliasing effects caused by downsampling during tokenization [24]. Chu et al. proposed conditional positional encoding that generates position embeddings adaptively based on input content rather than relying on fixed or purely learnable encodings [14]. Despite these advances, existing ViT-based approaches for pathological image classification still allocate computation uniformly across informative and uninformative regions, and they operate exclusively in the spatial domain, limiting their ability to capture diagnostically relevant frequency-domain features.

### *2.2 Lightweight Feature Extraction Techniques*

Reducing model parameters and computational cost while preserving performance has been a persistent research focus, particularly for deployment in resource-constrained clinical environments. MobileNet introduces depthwise separable convolutions to reduce operation counts dramatically [25]. ShuffleNet employs channel shuffle operations for efficient cross-group information exchange [26]. GhostNet proposes the Ghost module concept, which generates additional feature maps through cheap linear operations applied to a subset of intrinsic features, effectively halving the cost of standard convolution while maintaining representational capacity [27]. EfficientNet uses neural architecture search to jointly optimize network width, depth, and input resolution scaling [28].

Within the ViT architecture family, several works explore token-level efficiency. DynamicViT learns binary decisions for each token to predict and remove unimportant ones during inference [29]. EViT identifies inattentive tokens based on CLS attention scores and fuses them to preserve information while shortening the sequence [20]. A-ViT introduces adaptive token halting, allowing tokens to exit computation at different network depths depending on their complexity [30]. These methods demonstrate that significant savings can be achieved by exploiting information redundancy in token sequences. However, existing token pruning methods have not been tailored to the unique characteristics of histopathological images, where the distinction between informative and uninformative regions requires awareness of pathological semantics rather than simple visual saliency.

### *2.3 Frequency-Domain Analysis in Deep Learning*

Frequency-domain representations have shown considerable promise in various computer vision tasks. Xu et al. demonstrated that learning in the frequency domain can improve both accuracy and efficiency for image classification [17]. FcaNet introduces frequency channel attention by analyzing channel-wise information in the DCT frequency domain [31]. AFNO replaces standard self-attention with Fourier-domain token mixing, achieving global feature interaction at reduced cost [32]. In medical imaging, frequency-domain features are especially valuable for texture analysis. Histopathological textures, including chromatin patterns, cell membrane regularity, and stromal fiber orientation, exhibit distinctive frequency signatures that provide complementary information beyond spatial morphology [33]. Despite this potential, very few Vision Transformer architectures have systematically integrated frequency-domain analysis into their core attention mechanisms for pathological classification.

## 3. Method

### *3.1 Overall Architecture*

The overall architecture of GSDP-ViT is illustrated in Figure 1. The model adopts a single-branch design consisting of four sequential processing stages.

In the first stage, the input image $X \in \mathbb{R}^{H \times W \times 3}$ is processed through a multi-stage Ghost Convolutional Tokenizer (GCT) composed of cascaded Ghost modules, each followed by ReLU activation and max pooling. The GCT converts the raw image into a sequence of overlapping patch tokens $\mathbf{T}_0 \in \mathbb{R}^{N \times D}$, where N denotes the number of tokens and D denotes the embedding dimension. This stage extracts local features while reducing parameters relative to standard convolutional tokenization.

In the second stage, a learnable CLS token $\mathbf{t}_{cls} \in \mathbb{R}^{1 \times D}$ is prepended to the token sequence as a global representation carrier. Depthwise convolutional conditional position encoding (DwCPE) is applied to inject content-adaptive spatial position information, producing $\mathbf{T}_0^{pos} \in \mathbb{R}^{(N+1) \times D}$.

In the third stage, the positioned token sequence passes through L stacked Transformer encoder layers. Each layer contains Layer Normalization, SSDA, a residual connection, Layer Normalization, a Feed-Forward Network (FFN), another residual connection, and PTP. The computation in each encoder layer follows:

$$\mathbf{T}l' = SSDA(LN(\mathbf{T}l - 1)) + \mathbf{T}_{l-1}$$

$$\widehat{\mathbf{T}}_l = FFN(LN(\mathbf{T}_l')) + \mathbf{T}_l'$$

$$\mathbf{T}_l = PTP(\widehat{\mathbf{T}}_l)$$

where $l = 1,2,\ldots,L$. After PTP in each layer, the token count decreases: $|\mathbf{T}_l| < |\widehat{\mathbf{T}}_l|$, meaning computation progressively concentrates on the most relevant tokens.

In the fourth stage, Sequence Pooling aggregates information from all remaining tokens into a single representation vector after the final Transformer layer. This vector is then passed through a linear classification head to produce class prediction logits.

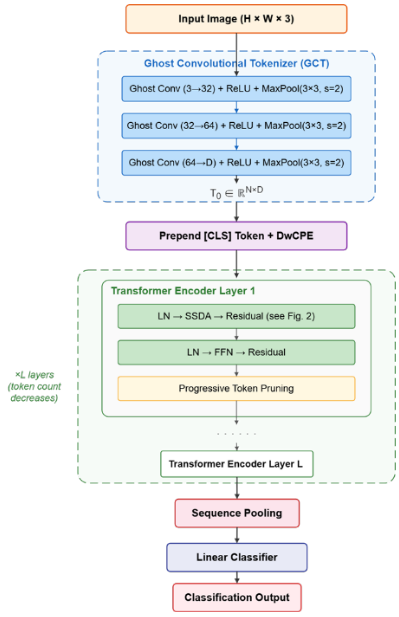

Input Image (H × W × 3)
Ghost Convolutional Tokenizer (GCT)
Ghost Conv (3→32) + ReLU + MaxPool(3×3, s=2)
Ghost Conv (32→64) + ReLU + MaxPool(3×3, s=2)
Ghost Conv (64→D) + ReLU + MaxPool(3×3, s=2)
$T_0 \in \mathbb{R}^{N\times D}$
Prepend [CLS] Token + DwCPE
Transformer Encoder Layer 1
LN → SSDA → Residual (see Fig. 2)
LN → FFN → Residual
Progressive Token Pruning
×L layers (token count decreases)
Transformer Encoder Layer L
Sequence Pooling
Linear Classifier
Classification Output

**Figure 1:** Overall architecture of the proposed GSDP-ViT model. The input image is first processed by a three-stage Ghost Convolutional Tokenizer (GCT) to generate patch tokens. After prepending a CLS token and adding depthwise conditional position encoding, the token sequence passes through L stacked Transformer encoder layers, each containing SSDA (detailed in Figure 2), FFN, and PTP. The token count decreases progressively across layers. Finally, Sequence Pooling aggregates the remaining tokens for linear classification.

### *3.2 Ghost Convolutional Tokenizer (GCT)*

The image patch embedding module in ViT models serves as a critical component that converts a two-dimensional image into a serialized one-dimensional token sequence for the Transformer encoder. Previous studies have demonstrated that replacing the linear projection used in the original ViT with convolution operations introduces a local receptive field, enables automatic learning of spatial pixel correlations, supports overlapping patch partitioning, and reduces the dependence on large-scale pretraining datasets.

However, standard convolutional tokenizers apply uniform convolution across all input channels, generating a substantial number of redundant feature maps. Research on trained neural networks has revealed that many output feature maps from a convolutional layer are similar to each other and can be treated as "ghosts" of a smaller set of intrinsic features [27]. In cervical histopathological images, this redundancy is particularly pronounced because large regions of normal epithelium, stroma, and background produce nearly identical feature responses across convolution channels.

Drawing on ideas from GhostNet [27], we propose the Ghost Convolutional Tokenizer (GCT), which replaces standard convolution in each tokenization stage with Ghost modules. For an input feature map $\mathbf{F}in \in \mathbb{R}^{H\prime\times W\prime\times Cin}$, a standard convolutional layer would generate $C_{out}$ feature maps using $C_{in} \times C_{out} \times k^2$ parameters. The Ghost module replaces this operation with a two-step process.

In the first step, standard convolution with $m = C_{out}/s$ kernels (where s is the Ghost ratio, set to 2) generates a compact set of intrinsic features:

$$\mathbf{F}_{int} = Convk \times k(\mathbf{F}_{in}) \in \mathbb{R}^{H'' \times W'' \times m}$$

In the second step, cheap depth wise convolution operations are applied to each intrinsic feature map to produce complementary ghost features:

$$\mathbf{F}_{ghost} = DWConvd \times d(\mathbf{F}_{int}) \in \mathbb{R}^{H'' \times W'' \times m}$$

The intrinsic and ghost features are then concatenated to form the final output:

$$\mathbf{F}_{out} = Concat(\mathbf{F}_{int}, \mathbf{F}_{ghost}) \in \mathbb{R}^{H'' \times W'' \times C_{out}}$$

This design reduces the total parameter count from $C_{in} \times C_{out} \times k^2$ to approximately $C_{in} \times m \times k^2 + m \times d^2$, achieving roughly 2× parameter reduction when s=2. The complete GCT consists of three cascaded stages:

$$\mathbf{E}_i = MaxPool(ReLU(Ghost(\mathbf{E}_{i-1})))$$

where $\mathbf{E}_0 = X$ is the input image. Overlapping max pooling with kernel size 3 and stride 2 captures edge information at patch boundaries. This tokenization approach offers two specific benefits for cervical histopathological images. On one hand, the intrinsic features capture the most discriminative pathological patterns such as nuclear atypia and mitotic figures, while ghost features capture complementary variations like chromatin gradient textures and cytoplasmic staining differences at minimal additional cost. On the other hand, using fewer learned parameters makes the tokenizer less susceptible to overfitting on the relatively small LDCH dataset.

### *3.3 Spectral–Spatial Dual Attention (SSDA)*

Standard self-attention in ViTs computes attention exclusively in the spatial domain, modeling inter-token relationships across different spatial patches. While this captures spatial arrangement patterns effectively, it does not explicitly access the frequency-domain characteristics within each token. In cervical histopathological images, critical diagnostic features manifest simultaneously in both domains. The spatial domain encodes lesion location, cellular arrangement, tissue architecture, and nuclear-to-cytoplasmic ratio. The frequency domain encodes chromatin texture coarseness, nuclear membrane smoothness, mitotic figure periodicity, and staining intensity gradients. To enable the model to reason jointly over both domains, we propose the Spectral–Spatial Dual Attention mechanism. The detailed architecture of SSDA is shown in Figure 2.

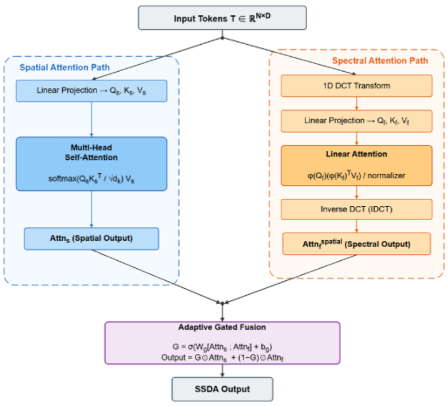

Input Tokens T ∈ ℝ^N×D
Spatial Attention Path
Linear Projection → Qs, Ks, Vs
Multi-Head Self-Attention
softmax(QsKs^T / √dk) Vs
Attns (Spatial Output)
Spectral Attention Path
1D DCT Transform
Linear Projection → Qf, Kf, Vf
Linear Attention
φ(Qf)(φ(Kf)^T Vf) / normalizer
Inverse DCT (IDCT)
Attnf^spatial (Spectral Output)
Adaptive Gated Fusion
G = σ(Wg[Attns ; Attnf] + bg)
Output = G⊙Attns + (1−G)⊙Attnf
SSDA Output

**Figure 2:** Architecture of the Spectral–Spatial Dual Attention (SSDA) module. Input tokens are processed through two parallel paths. The spatial path applies standard multi-head self-attention to capture structural

morphological patterns. The spectral path first transforms tokens to the frequency domain via 1D DCT, then performs lightweight linear attention on frequency components, and finally applies inverse DCT to return to the spatial domain. A learnable gating mechanism adaptively fuses the outputs of the two paths.

### *3.3.1 Spatial Attention Path*

The spatial attention path follows the standard multi-head self-attention formulation. Given input tokens $\mathbf{T} \in \mathbb{R}^{N \times D}$, three linear projections produce queries, keys, and values:

$$\mathbf{Q}_s = \mathbf{T}\mathbf{W}_Q^s, \quad \mathbf{K}_s = \mathbf{T}\mathbf{W}_K^s, \quad \mathbf{V}_s = \mathbf{T}\mathbf{W}_V^s$$

The spatial attention output is:

$$Attn_s = softmax\left(\frac{\mathbf{Q}_s\mathbf{K}_s^\top}{\sqrt{d_k}}\right)\mathbf{V}_s$$

where $d_k = D/h$ and h is the number of attention heads. This path preserves the standard ViT capability to model global spatial relationships among all patch tokens.

### *3.3.2 Spectral Attention Path*

The spectral attention path first transforms each token from the spatial domain to the frequency domain using the 1D Discrete Cosine Transform (DCT). For each token $\mathbf{t}_i \in \mathbb{R}^D$, the DCT is applied along the embedding dimension:

$$\tilde{\mathbf{t}}_i[k] = \sum n = 0^{D-1} \mathbf{t}_i[n] \cos\left[\frac{\pi}{D}\left(n + \frac{1}{2}\right)k\right], \quad k = 0,1,\dots,D-1$$

This operation produces spectral tokens $\tilde{\mathbf{T}} \in \mathbb{R}^{N \times D}$, where each element represents a frequency component rather than a spatial feature. Separate linear projections produce spectral queries, keys, and values:

$$\mathbf{Q}_f = \tilde{\mathbf{T}}\mathbf{W}_Q^f, \quad \mathbf{K}_f = \tilde{\mathbf{T}}\mathbf{W}_K^f, \quad \mathbf{V}_f = \tilde{\mathbf{T}}\mathbf{W}_V^f$$

To keep the computational cost of the spectral path manageable, we adopt a kernel-based linear attention formulation rather than standard SoftMax attention:

$$Attn_f = \frac{\phi(\mathbf{Q}_f)\left(\phi(\mathbf{K}_f)^\top\mathbf{V}_f\right)}{\phi(\mathbf{Q}_f) \cdot \phi(\mathbf{K}_f)^\top\mathbf{1}}$$

where $\phi(\cdot) = ELU(\cdot) + 1$ serves as a non-negative feature mapping function. By computing the matrix product $\phi(\mathbf{K}_f)^\top\mathbf{V}_f$ first (yielding a $d_k \times d_k$ matrix), the complexity is reduced from $O(N^2 d_k)$ to $O(N d_k^2)$. After spectral attention computation, the inverse DCT (IDCT) transforms the result back to the spatial domain:

$$Attn_f^{spatial} = IDCT(Attn_f)$$

### *3.3.3 Adaptive Gated Fusion*

The spatial and spectral path outputs are combined through a learnable gating mechanism that determines the relative contribution of each path for every token and feature dimension:

$$\mathbf{G} = \sigma\left(\mathbf{W}_g[Attn_s;;Attn_f^{spatial}] + \mathbf{b}_g\right)$$

$$SSDA(\mathbf{T}) = \mathbf{G} \odot Attn_s + (1 - \mathbf{G}) \odot Attn_f^{spatial}$$

Here $\sigma(\cdot)$ is the sigmoid function, $[\cdot;\cdot]$ denotes concatenation along the feature dimension, $\mathbf{W}_g \in \mathbb{R}^{D\times 2D}$ and $\mathbf{b}_g \in \mathbb{R}^D$ are learnable parameters, and $\odot$ denotes element-wise multiplication. Through this mechanism, the model learns to assign higher spectral weights to texture-rich regions where frequency patterns carry critical diagnostic significance, and higher spatial weights to structurally defined regions where global layout matters more.

### *3.4 Progressive Token Pruning (PTP)*

Cervical histopathological images typically contain large areas of diagnostically irrelevant content, including background, normal epithelial tissue, and stromal areas, alongside relatively small but critical lesion regions. Processing all tokens uniformly through the entire network wastes computation on uninformative regions and may inject noise that harms classification. The Progressive Token Pruning strategy addresses this by dynamically identifying and removing the least informative tokens after each Transformer layer, progressively focusing resources on pathologically relevant regions.

Token Importance Scoring. After each encoder layer l, an importance score$s_i$ is computed for each patch token $\mathbf{t}_i$ (excluding the CLS token) by averaging the CLS-to-token attention weights across all heads:

$$s_i^{(l)} = \frac{1}{h}\sum_{j=1}^{h} A_{0,i}^{(j)}$$

Where $A_{0,i}^{(j)}$ denotes the attention weight from the CLS token to token i in head j. Tokens receiving higher CLS attention are considered more diagnostically important, since the CLS token aggregates global information for the final classification decision.

Pruning with Information Preservation. At each layer l, a retention ratio$\rho_l$ determines the fraction of tokens to keep. Tokens are ranked by importance, and the top $\lceil \rho_l \cdot N_l \rceil$ tokens (plus the CLS token, which is always retained) form the input to the next layer. To prevent abrupt information loss, a weighted aggregation of the pruned tokens is added to the CLS token:

$$\mathbf{t}_{cls}^{updated} = \mathbf{t}_{cls} + \alpha \sum_{i\in\mathcal{P}} \bar{s}_{\ i} \cdot \mathbf{t}_i$$

where $\mathcal{P}$ is the set of pruned token indices, $\bar{s}_i = s_i / \sum j \in \mathcal{P} s_j$ are normalized importance scores among pruned tokens, and $\alpha$ is a learnable scalar initialized to 0.1. This mechanism preserves information from less important but potentially useful regions in compressed form within the CLS token.

Progressive Schedule. Retention ratios follow a linear decay across layers:

$$\rho_l = 1 - \frac{l}{L} \cdot (1 - \rho_{\min})$$

where L is the total number of layers and$\rho_{\min}$ is the minimum retention ratio set to 0.5. For a 4-layer network, the resulting schedule is: Layer 1 retains 88%, Layer 2 retains 75%, Layer 3 retains 63%, and Layer 4 retains 50% of tokens. This progressive design ensures that early layers process the full spatial information for comprehensive feature extraction, while deeper layers increasingly concentrate computation on the most diagnostically relevant tokens.

Training Loss. A diversity regularization term is added to the standard cross-entropy loss to prevent degenerate pruning patterns where the model consistently focuses on the same small region:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \left( -\frac{1}{L} \sum l = 1^{L} H(\mathbf{s}^{(l)}) \right)$$

where $H(\cdot)$ denotes the entropy of the normalized importance score distribution and $\lambda = 0.1$ is a balancing coefficient. This loss term encourages higher-entropy attention distributions, promoting comprehensive coverage of diverse pathological regions across the image.

### *3.5 Position Encoding and Classification Head*

For position encoding, we adopt depth wise convolutional conditional position encoding (DwCPE), which generates content-adaptive position information through 2D depth wise convolution. The token sequence is reshaped into a 2D feature map according to the spatial arrangement of patches, processed by a $3 \times 3$ depthwise convolution with zero padding, and flattened back to a 1D sequence. This encoding introduces spatial locality bias compatible with the 2D structure of histopathological images, adapts to varying input resolutions, and adds only 9D parameters.

For the classification head, after the final Transformer layer, Sequence Pooling computes learned attention weights over all remaining tokens and produces a weighted aggregation. This aggregated vector is passed through a linear layer to generate classification logits for the nine cervical lesion categories.

## 4. Experiments and Results

### *4.1 Dataset and Implementation Details*

LDCH Dataset. Experiments are conducted on the LDCH dataset collected by the Landing Artificial Intelligence Pathology Diagnostic Center. The dataset contains 286 cervical histopathological images scanned at 40× magnification using a Jiangfeng Biotechnology scanner. Each image is accompanied by local annotations marking multiple lesion regions, which are divided into nine categories: CIN I (1048 regions), CIN II (1669 regions), CIN III (302 regions), CIN II involving glands (4494 regions), CIN III involving glands (125 regions), Glandular tissue (8248 regions), Adenosquamous metaplasia (322 regions), Koilocytes (69 regions), and Squamous epithelium (5913 regions). Each annotation was reviewed and confirmed by at least two associate chief physician-level pathologists. The dataset exhibits severe class imbalance, with the largest category (Glandular tissue) containing approximately 120 times more samples than the smallest (Koilocytes). Data was randomly divided into training, validation, and test sets at a 6:3:1 ratio.

Implementation Details. The model was implemented in PyTorch and trained on a computing platform equipped with a Hygon Z100 16GB deep learning accelerator (DCU). The AdamW optimizer was used with an initial learning rate of $1 \times 10^{-3}$ and weight decay of $1 \times 10^{-4}$. A cosine annealing learning rate schedule was applied over 100 epochs with 20 epochs of linear warm-up. The batch size was 32. Standard data augmentation including random horizontal and vertical flips, random rotation, color jitter, and random erasing was applied during training. Each experiment was repeated five times, and mean values are reported.

Model Configuration. The default configuration uses embedding dimension D = 128, L = 4 Transformer encoder layers, h = 4 attention heads, FFN expansion ratio of 2, Ghost ratio s = 2, minimum token retention ratio $\rho_{\min} = 0.5$, and diversity loss weight $\lambda = 0.1$.

### *4.2 Evaluation Metrics*

Following established evaluation protocols, we adopt the following metrics. Accuracy (ACC) measures the proportion of correctly classified samples. Macro-F1 Score is the arithmetic mean of per-class F1 scores, treating each class equally regardless of sample count. This metric is especially important given the severe class imbalance in the LDCH dataset and serves as a key indicator of how well the model handles minority classes. Recall and Precision are reported as macro-averaged values across all classes. Parameters (M) represents total learnable parameters in millions.

### *4.3 Comparison with State-of-the-Art Methods*

Table 1 Presents the Comparison Between GSDP-ViT and Various Existing Methods on the LDCH Dataset.

**Table 1:** Comparison of Experimental Results Between GSDP-ViT and Other Methods on the LDCH Dataset**.**

| Method | ACC (%) | F1 (%) | Recall (%) | Precision (%) | Params (M) |
|---|---|---|---|---|---|
| MobileNet-v2 | 81.21 | 46.56 | 44.17 | 56.84 | 2.23 |
| ShuffleNet-v2 | 80.98 | 45.85 | 42.72 | 57.96 | 0.35 |
| ResNet | 81.92 | 60.04 | 59.35 | 40.57 | 1.16 |
| EfficientFormer | 80.83 | 42.23 | 40.20 | 60.83 | 11.39 |
| T2T-ViT | 62.93 | 23.53 | 25.52 | — | 4.00 |
| CeiT | 82.69 | 57.40 | 54.08 | 68.97 | 6.16 |
| TinyViT | 71.24 | 25.75 | 28.61 | 40.57 | 5.07 |
| DGTNet | 84.35 | — | — | — | — |
| DB-BDGANet | 86.61 | 65.67 | 62.53 | — | — |
| EVT | 85.52 | 62.13 | 58.09 | 72.49 | 0.53 |
| PCPECT | 87.38 | 67.36 | 64.00 | 74.19 | 0.62 |
| GSDP-ViT (ours) | 87.85 | 68.21 | 65.27 | 73.56 | 0.47 |

The results reveal several noteworthy findings. GSDP-ViT achieves the highest accuracy of 87.85%, exceeding the recent dual-branch PCPECT model (87.38%) and the EVT baseline (85.52%) by 0.47% and 2.33% respectively. This confirms the effectiveness of the proposed architecture for cervical histopathological classification.

The macro-F1 score of 68.21% is the highest among all compared methods. This suggests that GSDP-ViT handles the severe class imbalance in the LDCH dataset more effectively than competing approaches. The improvement over PCPECT (67.36%) can be attributed to the SSDA mechanism, which captures frequency-domain texture features critical for distinguishing morphologically similar lesion subtypes.

With only 0.47M parameters, GSDP-ViT is the most parameter-efficient model among those achieving competitive accuracy. It uses 24% fewer parameters than PCPECT (0.62M), 11% fewer than EVT (0.53M), and far fewer than models such as Efficient Former (11.39M) and CeiT (6.16M). This efficiency stems primarily from the Ghost Convolutional Tokenizer.

Comparing across model families, pure CNN models (MobileNet-v2, ShuffleNet-v2, ResNet) achieve moderate accuracy but low F1 scores due to limited receptive fields. Pure ViT models (T2T-ViT, TinyViT) perform poorly on this small dataset because they lack inductive biases suited to limited training data. Hybrid and enhanced approaches (CeiT, EVT, PCPECT, GSDP-ViT) consistently outperform both pure paradigms. Notably, GSDP-ViT surpasses the dual-branch PCPECT and

DB-BDGANet using a simpler single-branch architecture, demonstrating that careful module design can substitute for architectural complexity.

### 4.4 Ablation Experiments

To evaluate each component's contribution, ablation experiments were conducted by progressively adding modules to the EVT baseline. Results are shown in Table 2.

**Table 2:** Ablation Study Results on the LDCH Dataset.

| Configuration | ACC (%) | F1 (%) | Recall (%) | Precision (%) | Params (M) |
|---|---|---|---|---|---|
| Baseline (EVT) | 85.52 | 62.13 | 58.09 | 72.49 | 0.53 |
| + GCT | 85.89 | 63.05 | 59.41 | 71.82 | 0.31 |
| + SSDA | 86.47 | 65.83 | 62.14 | 73.28 | 0.58 |
| + PTP | 86.15 | 63.92 | 60.28 | 72.91 | 0.53 |
| + GCT + SSDA | 86.91 | 66.52 | 63.05 | 73.41 | 0.36 |
| + GCT + PTP | 86.44 | 64.18 | 60.79 | 72.56 | 0.31 |
| + SSDA + PTP | 87.03 | 66.89 | 63.68 | 73.12 | 0.57 |
| Full model (GCT+SSDA+PTP) | 87.85 | 68.21 | 65.27 | 73.56 | 0.47 |

The ablation results yield several observations. GCT alone improves accuracy by 0.37% while reducing parameters by 41.5% (from 0.53M to 0.31M). This confirms that Ghost modules effectively maintain feature diversity while eliminating redundancy and easing overfitting risk.

SSDA alone contributes the largest individual improvement, with a 0.95% accuracy gain and a 3.70% F1 increase. This substantial F1 improvement validates that spectral-domain attention captures frequency-level diagnostic features that spatial attention alone misses, which is particularly helpful for distinguishing minority classes with subtle morphological differences.

PTP alone improves accuracy by 0.63% and F1 by 1.79% without adding parameters. It demonstrates effectiveness in reducing noise from uninformative regions and directing the model toward pathologically relevant areas.

The full combination achieves the best performance, with improvements exceeding the sum of individual contributions. This reveals important synergistic effects: GCT produces cleaner initial tokens that enable SSDA to compute more meaningful spectral features, while PTP removes tokens that would otherwise dilute spectral attention with irrelevant frequency information.

### 4.5 Sensitivity Analysis

Additional experiments were performed to evaluate sensitivity to key hyperparameters. The results are summarized in Table 3.

**Table 3:** Sensitivity Analysis of Ghost Ratio and Minimum Token Retention Ratio.

| Ghost Ratio s | ACC (%) | F1 (%) | Params (M) |
|---|---|---|---|
| 1 (standard conv) | 87.52 | 67.58 | 0.58 |
| 2 | 87.85 | 68.21 | 0.47 |
| 4 | 87.03 | 66.42 | 0.42 |
| $\rho_{\min}$ | **ACC (%)** | **F1 (%)** | **GFLOPs** |

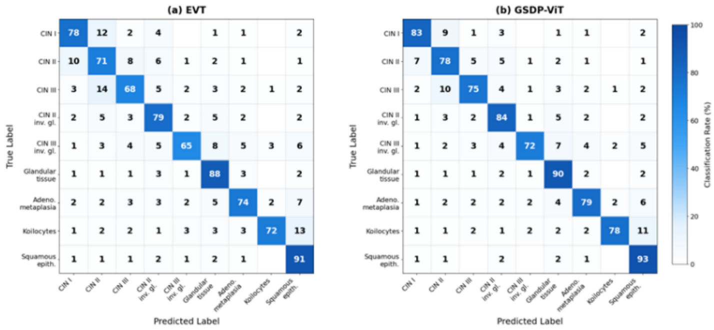
| | | | |
|---|---|---|---|
| 0.3 | 86.91 | 66.73 | 0.73 |
| 0.5 | 87.85 | 68.21 | 0.89 |
| 0.7 | 87.44 | 67.52 | 1.02 |
| 1.0 (no pruning) | 87.31 | 67.15 | 1.15 |

For the Ghost ratio, s=2 provides the best balance between parameter reduction and feature quality. Setting s=1 (equivalent to standard convolution) increases parameters without meaningful accuracy gains. Setting s=4 reduces parameters further but degrades feature quality, leading to a 0.82% accuracy drop.

For the minimum retention ratio, $\rho_{\min} = 0.5$ yields optimal results. More aggressive pruning causes excessive information loss, reducing accuracy by 0.94%. Less aggressive pruning retains too many uninformative tokens, increasing computation without improving and even slightly reducing performance. Interestingly, no pruning at all) performs slightly worse than moderate pruning, suggesting that removing uninformative tokens actively helps by eliminating noisy features rather than merely saving computation.

### 4.6 Visualization Analysis

To qualitatively assess model behavior, we present confusion matrices comparing the EVT baseline and the proposed GSDP-ViT in Figure 3. Both matrices are computed on the LDCH test set.



(a) EVT

| | CIN I | CIN II | CIN III | CIN II inv. gl. | CIN III inv. gl. | Glandular tissue | Adeno. metaplasia | Koilocytes | Squamous epith. |
|---|---|---|---|---|---|---|---|---|---|
| CIN I | 78 | 12 | 2 | 4 | | 1 | 1 | | 2 |
| CIN II | 10 | 71 | 8 | 6 | 1 | 2 | 1 | | 1 |
| CIN III | 3 | 14 | 68 | 5 | 2 | 3 | 2 | 1 | 2 |
| CIN II inv. gl. | 2 | 5 | 3 | 79 | 2 | 5 | 2 | | 2 |
| CIN III inv. gl. | 1 | 3 | 4 | 5 | 65 | 8 | 5 | 3 | 6 |
| Glandular tissue | 1 | 1 | 1 | 3 | 1 | 88 | 3 | | 2 |
| Adeno. metaplasia | 2 | 2 | 3 | 3 | 2 | 5 | 74 | 2 | 7 |
| Koilocytes | 1 | 2 | 2 | 1 | 3 | 3 | 3 | 72 | 13 |
| Squamous epith. | 1 | 1 | 1 | 2 | 1 | 2 | 1 | | 91 |

(b) GSDP-ViT

| | CIN I | CIN II | CIN III | CIN II inv. gl. | CIN III inv. gl. | Glandular tissue | Adeno. metaplasia | Koilocytes | Squamous epith. |
|---|---|---|---|---|---|---|---|---|---|
| CIN I | 83 | 9 | 1 | 3 | | 1 | 1 | | 2 |
| CIN II | 7 | 78 | 5 | 5 | 1 | 2 | 1 | | 1 |
| CIN III | 2 | 10 | 75 | 4 | 1 | 3 | 2 | 1 | 2 |
| CIN II inv. gl. | 1 | 3 | 2 | 84 | 1 | 5 | 2 | | 2 |
| CIN III inv. gl. | 1 | 2 | 3 | 4 | 72 | 7 | 4 | 2 | 5 |
| Glandular tissue | 1 | 1 | 1 | 2 | 1 | 90 | 2 | | 2 |
| Adeno. metaplasia | 1 | 2 | 2 | 2 | 2 | 4 | 79 | 2 | 6 |
| Koilocytes | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 78 | 11 |
| Squamous epith. | 1 | 1 | | 2 | | 2 | 1 | | 93 |

**Figure 3:** Confusion matrices on the LDCH test set for (a) the EVT baseline and (b) the proposed GSDP-ViT. GSDP-ViT demonstrates notable improvements in distinguishing difficult category pairs, particularly reducing CIN II versus CIN III confusion, improving CIN III involving glands recognition, and enhancing Koilocytes identification despite the extremely small sample size of that category.

The confusion matrices reveal several important improvements achieved by GSDP-ViT. The misclassification rate between CIN II and CIN III, which represents one of the most challenging distinctions in cervical pathology, decreases noticeably. In the baseline model, 14% of CIN III samples are misclassified as CIN II, whereas this rate drops to 10% in GSDP-ViT. This improvement can be attributed to the SSDA mechanism, which captures chromatin texture frequency differences that help distinguish the two dysplasia grades.

The recognition rate for CIN III involving glands improves from 65% to 72%. This is the category with the second-smallest sample count (125 regions), and accurate identification is clinically important because glandular involvement indicates a higher risk of progression. The spectral attention path appears to detect disruption patterns in glandular architecture that manifest as distinctive frequency signatures.

Koilocytes identification improves from 72% to 78% despite this category having the smallest sample size (69 regions). Koilocytes are characterized by a distinctive perinuclear clearing pattern, and the PTP mechanism likely helps by concentrating attention on these small but diagnostically distinctive cellular features rather than spreading computation across large uninformative areas.

Squamous epithelium, the second-largest category, shows improvement from 91% to 93%, and Adenosquamous metaplasia improves from 74% to 79%. These results demonstrate consistent performance gains across both majority and minority classes, further supporting the robustness of the proposed approach.

## 5. Conclusion

This paper presents GSDP-ViT, a lightweight Vision Transformer designed for cervical histopathological image classification. The model incorporates three modules that work in concert to address the specific challenges of this task. The Ghost Convolutional Tokenizer achieves parameter-efficient feature extraction by augmenting a compact set of intrinsic features through cheap depthwise operations, cutting the tokenizer parameter count by roughly half while preserving feature diversity. The Spectral–Spatial Dual Attention mechanism extends standard self-attention with a parallel frequency-domain path that transforms tokens via DCT, enabling joint processing of spatial morphological features and frequency-domain textural signatures through an adaptive gating mechanism. The Progressive Token Pruning strategy dynamically reallocates computational resources toward diagnostically relevant regions by gradually removing uninformative tokens at each layer, simultaneously reducing cost and improving classification accuracy through noise reduction.

On the LDCH dataset, GSDP-ViT achieves a classification accuracy of 87.85% and a macro-F1 score of 68.21% with only 0.47 million parameters and 0.89 GFLOPs, outperforming all compared methods including CNN models, ViT variants, and recent hybrid and dual-branch architectures. Ablation studies confirm each component's effectiveness and reveal synergistic interactions that produce combined improvements exceeding the sum of individual contributions.

The results suggest that effective cervical histopathological image classification does not necessarily require complex multi-branch architectures or large parameter budgets. Thoughtful integration of frequency-domain awareness, efficient feature generation, and adaptive computation allocation can deliver strong performance within a compact single-branch framework. The model offers a practical and deployable solution for computer-aided cervical cancer screening, particularly in clinical environments where both accuracy and computational efficiency are essential. Future work will explore extending the framework to whole-slide image analysis and incorporating additional frequency-domain transforms such as wavelet decomposition for multi-scale spectral feature extraction.

## References

[1] Arbyn M, Weiderpass E, Bruni L, et al. Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis. The Lancet Global Health, 2020, 8(2): e191–e203.

[2] Schiffman M, Doorbar J, Wentzensen N, et al. Carcinogenic human papillomavirus infection. Nature Reviews Disease Primers, 2016, 2: 16086.

[3] Arbyn M, Anttila A, Jordan J, et al. European guidelines for quality assurance in cervical cancer screening. Annals of Oncology, 2010, 21(3): 448–458.

[4] Gurcan M N, Boucheron L E, Can A, et al. Histopathological image analysis: A review. IEEE Reviews in

Biomedical Engineering, 2009, 2: 147–171.

[5] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition.

[8] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention.

[9] Lin J, Roy S, Li H, et al. Super vision transformer.

[10] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks.

[11] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis. Medical Image Analysis, 2017, 42: 60–88.

[12] Chen R J, Lu M Y, Wang J, et al. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis. IEEE Transactions on Medical Imaging, 2021, 40(3): 757–770.

[13] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows.

[14] Chu X, Tian Z, Wang Y, et al. Conditional positional encodings for vision transformers.

[15] Guo J, Han K, Wu H, et al. CMT: Convolutional neural networks meet vision transformers.

[16] Dong X, Bao J, Chen D, et al. CSWin Transformer: A general vision transformer backbone with cross-shaped windows.

[17] Xu K, Qin M, Sun F, et al. Learning in the frequency domain.

[18] Han K, Xiao A, Wu E, et al. Transformer in transformer.

[19] Yuan L, Chen Y, Wang T, et al. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet.

[20] Liang Y, Ge C, Tong Z, et al. Not all tokens are equal: Human-centric visual analysis via token reorganization[C]//CVPR. 2022.

[21] Dai Y, Gao Y, Liu F. TransMed: Transformers advance multi-modal medical image classification. Diagnostics, 2021, 11(8): 1384.

[22] Yu S, Li J, Liu Z, et al. MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification.

[23] Hassani A, Walton S, Li J, et al. Escaping the big data paradigm with compact transformers.

[24] Yao H, Zhang X, Zhou X, et al. EVT: Enhanced Vision Transformer with wavelet position embedding for histopathological image classification. 2024.

[25] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks.

[26] Ma N, Zhang X, Zheng H, et al. ShuffleNet V2: Practical guidelines for efficient CNN architecture design.

[27] Han K, Wang Y, Tian Q, et al. GhostNet: More features from cheap operations.

[28] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks.

[29] Rao Y, Zhao W, Liu B, et al. DynamicViT: Efficient vision transformers with dynamic token sparsification.

[30] Yin H, Molchanov P, Alvarez J M, et al. A-ViT: Adaptive tokens for efficient vision transformer.

[31] Qin Z, Zhang P, Wu F, et al. FcaNet: Frequency channel attention networks.

[32] Guibas J, Mardani M, Karras T, et al. Adaptive Fourier neural operators: Efficient token mixers for transformers.

[33] Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Computational and Structural Biotechnology Journal, 2018, 16: 34–42.