Precise cervical cancer cell boundary denoising and segmentation with adaptive wavelet-spectral enhancement

Lalasa Mukku [a,1,*], Manjunath Ramanna Lamani [b,2], Lavanya Hegde [c,3], Prathima Mahapurush [d,4], Shivanandaswamy Mahapurush [b,5]

[a] CHRIST (Deemed to be University), Bangalore 560074, India
[b] Moodlakatte Institute of Technology, Kundapura 576217, India
[c] GOVERNMENT SKSJTI, Bangalore 560001, India
[d] SKSVMACET Gadag, India
[1] mlalasa2020@gmail.com; [2] manjunathlamani01@gmail.com ; [3] drlavanyahegde@gmail.com; [4] pmahapurush@gmail.com;
[5] shivu1201@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received October 27, 2025
Revised December 22, 2025
Accepted February 7, 2026
Available online February 28, 2026

Keywords
Adaptive wavelet-spectral
Cervical cells
Cervical cancer
Cell boundary denoising
Segmentation enhancement

ABSTRACT

Accurate segmentation of cell nuclei in cervical cytology images is crucial for automated cervical cancer screening, yet existing methods struggle with blurred boundaries, noise-induced degradation, and topologically implausible predictions. The current research proposes Cell-Seg Tool, a novel triplet-branch diffusion AI tool that synergistically integrates three innovations to address these limitations. The Wavelet-Enhanced Contour Refinement Branch employs a learnable multi-scale discrete wavelet transform with adaptive coefficient attention to dynamically enhance boundary features across horizontal, vertical, and diagonal orientations. The Adaptive Spectral Noise Suppression module performs dual-domain processing using DCT-based filtering and uncertainty-guided fusion, coupled with bidirectional anchor semantic feedback to couple cross-branch information. The Topology-Aware Hybrid Loss integrates a focal Tversky loss, a persistent homology loss, a directional boundary loss, a skeleton completeness loss, and a diffusion-noise MSE loss for multi-objective optimization. Comprehensive experiments on multiple datasets demonstrate superior performance, achieving 94.45% Dice coefficient and 19.2% reduction in boundary localization error compared to state-of-the-art methods. Unlike prior work that applies these techniques independently, this work demonstrates that their adaptive, synergistic integration within a diffusion-based framework yields substantial improvements in boundary accuracy and topological correctness.

1. Introduction

Cervical cancer is the fourth most common cancer among women worldwide. According to the latest statistics from the World Health Organization (WHO), approximately 604,000 new cases and 342,000 deaths were reported globally in 2020, with the vast majority occurring in low- and middle-income countries [1], [2]. The disease progression from precancerous cervical intraepithelial neoplasia (CIN) to invasive carcinoma typically spans several years, offering a critical window for early detection and intervention. Papanicolaou (Pap) smear cytology and liquid-based cytology remain the gold standard screening modalities, enabling identification of abnormal cells through microscopic examination [3] of exfoliated cervical specimens [4]. However, manual cytological screening is labor-intensive, requires highly trained cytopathologists, and suffers from considerable inter-observer variability, factors that collectively limit screening coverage and diagnostic consistency, particularly in resource-constrained

settings [5]. Accurate segmentation of cell nuclei in these cytological images is fundamental to automated screening systems, as nuclear morphology, chromatin distribution, and nuclear-to-cytoplasmic (N/C) ratio serve as primary diagnostic indicators for distinguishing normal cells from dysplastic and malignant cells [6], [7].

Diffusion probabilistic models (DPMs) represent a fundamentally different generative paradigm that has recently demonstrated remarkable capabilities in image synthesis and reconstruction tasks. Unlike discriminative models [8], [9], DPMs learn to model the joint distribution p(image, mask) through a two-stage process: a forward diffusion process that progressively corrupts data by adding Gaussian noise [10] according to a predefined schedule, and a reverse denoising process that learns to iteratively reconstruct clean data from pure noise. This generative formulation offers several theoretical and practical advantages for segmentation tasks. First, by modelling the complete data distribution rather than just decision boundaries, DPMs can better handle ambiguous regions where multiple plausible segmentations exist, a common scenario in medical images with blurred boundaries or overlapping structures. Second, the iterative refinement process inherent to reverse diffusion enables progressive improvement of segmentation quality through multiple denoising steps, analogous to how human experts refine their interpretations through careful examination. Third, the probabilistic nature of DPMs naturally accommodates uncertainty quantification [11], potentially providing confidence estimates valuable for clinical decision-making.

Recent works have begun exploring diffusion models for medical image segmentation, demonstrating competitive or superior performance compared to discriminative approaches. MedSegDiff [12] introduced dynamic conditional encoding to establish state-adaptive conditions for each sampling step and incorporated frequency-domain filtering (FF-Parser) to suppress high-frequency noise along skip connections. BerDiff [13] proposed using Bernoulli noise rather than Gaussian noise to better match the discrete, binary nature of segmentation masks. EnsDiff [14] leveraged ensemble predictions through multiple stochastic samplings to reduce prediction variance and improve robustness. AmDiff [15] focused on modelling segmentation ambiguity by generating multiple plausible outputs for uncertain regions. While these pioneering efforts have validated the potential of diffusion models for segmentation tasks, critical challenges remain inadequately addressed.

The fundamental challenge lies in the inherent tension between the noise injection required for diffusion-based generation and the need for precise localization in segmentation tasks. During the forward diffusion process, progressive noise addition corrupts not only the target masks but also degrades feature representations extracted by the encoder network. This noise propagates through intermediate layers and can lead to semantic confusion, where the model struggles to distinguish genuine structural features from noise-induced artefacts, particularly in cervical cytology, where nuclear boundaries often exhibit low contrast. Existing noise mitigation strategies, such as MedSegDiff's fixed Fourier filtering, apply uniform frequency cutoffs that cannot adapt to spatially varying noise characteristics or preserve important high-frequency details corresponding to fine boundaries. Moreover, the lack of explicit boundary modelling in current diffusion-based methods means that edge information must be implicitly learned through the denoising objective, which provides insufficient supervision for precise boundary localization in challenging cases with weak gradients or overlapping structures.

Beyond the noise and boundary challenges, topological correctness represents another critical dimension largely overlooked in existing diffusion-based segmentation approaches. In cervical cytology, nuclei should exhibit specific topological properties: each cell typically contains a single connected nuclear region without holes or fragmentation. Conventional loss functions, such as Dice loss or cross-entropy, optimize for pixel-wise accuracy and region overlap but provide no explicit constraints on topological structure. Consequently, models can produce anatomically implausible results, such as splitting single nuclei into multiple disconnected components or introducing spurious holes within nuclear regions that violate biological constraints and can lead to incorrect downstream analyses. While some recent works have explored topological losses based on persistent homology, their integration with diffusion models for medical image segmentation remains largely unexplored.

Despite significant advances in deep learning-based segmentation, fundamental challenges remain inadequately addressed in cervical nucleus segmentation. Boundary ambiguity: Nuclear boundaries in cervical cytology images frequently exhibit severe ambiguity. In addition, conventional loss functions used in medical image segmentation primarily focus on pixel-wise accuracy and boundary localization, while failing to enforce topological constraints on predicted segmentation masks. In cervical cytology, where nuclei exhibit complex morphologies including bean-shaped, lobulated, and irregular contours, maintaining topological correctness is essential for accurate lesion characterization.

To address these fundamental limitations, this paper proposes a comprehensive tool that synergistically integrates three key innovations: learnable wavelet-based boundary enhancement, adaptive frequency-domain noise suppression with bidirectional semantic coupling, and topology-aware multi-objective optimization. The core insight underlying our approach is that achieving high-quality segmentation in challenging scenarios requires explicit modelling and optimization of multiple complementary aspects, including semantic regions, boundary structures, noise robustness, and topological correctness, rather than relying solely on end-to-end learning with a single objective function. The main contributions of this manuscript are as follows:

- Wavelet-Enhanced Contour Refinement Branch. A fully learnable boundary enhancement module that employs multi-scale 2D Discrete Wavelet Transform with Adaptive Wavelet Coefficient Attention (AWCA) to dynamically weight horizontal, vertical, and diagonal edge components, generating Contour Saliency Maps that are progressively integrated into both diffusion and semantic branches through gated Multi-level Contour Feature Integration (MCFI) modules for robust boundary localization. The primary novelty of this work lies not in the individual components, but in their coordinated design and interaction within a triplet-branch diffusion architecture.
- Adaptive Spectral Noise Suppression (ASNS) Module. A novel dual-domain denoising mechanism featuring parallel spatial-spectral processing with DCT-based adaptive filtering, uncertainty-guided fusion using local variance estimation, and bidirectional anchor semantic feedback that establishes cross-branch coupling to enhance feature quality while preserving structural details during diffusion.
- Topology-Aware Hybrid Loss (TAHL). Multi-component supervision incorporating five complementary objectives (focal Tversky, persistent homology via Ripser library, Scharr-based directional gradients, morphological skeleton matching, diffusion noise MSE) with systematic grid-searched weights (λ_{tversky} 2.0, $\lambda_{\text{topology}}$ 0.5, $\lambda_{\text{direction}} = 0.3$, $\lambda_{\text{skeleton}} = 0.4$) ensuring anatomically plausible segmentation with correct topology and complete contours.

The rest of the manuscript is organized as follows: Section 2 reviews the related literature, Section 3 presents the methodology, Section 4 discusses the results, and Section 5 concludes the study.

2. Related Work

Automated segmentation of cervical cell nuclei has been pursued through diverse methodological approaches, ranging from traditional unsupervised clustering techniques to contemporary deep learning architectures. Early methods predominantly employed intensity-based thresholding, morphological operations, and geometric shape modelling to extract nuclear regions, while recent advances have leveraged CNNs and attention mechanisms to capture complex morphological patterns [16]–[19]. Most recently, DPMs have emerged as a powerful generative paradigm for medical image segmentation, offering superior handling of ambiguous regions through iterative denoising processes. However, persistent challenges remain in accurately localizing blurred boundaries, suppressing noise-induced feature degradation, and ensuring topologically plausible predictions, limitations that motivate the development of our proposed AI tool. This section reviews relevant literature across traditional segmentation methods, deep learning approaches, diffusion-based models, and boundary-aware techniques, highlighting their contributions and identifying gaps that our work addresses.

The traditional cell segmentation methods include cluster models such as k-means. For instance, [20] developed a new method, Adaptive Nucleus Shape Modeling, that segments cell nuclei using multilevel thresholding and fits ellipses to their shapes. The research uses multi-level thresholding to identify nuclear regions while accounting for variability in nuclear morphology. Extracts texture features from the inscribed rectangle within the fitted ellipse; these features characterize nuclei's texture and shape at the nucleus level. Research by [21] attempted segmentation of cervical cell images using an unsupervised machine learning technique. A multi-scale hierarchical segmentation algorithm to partition these regions based on homogeneity and circularity. They tested a two-fold method that begins with Nucleus and cytoplasm segmentation, followed by Hierarchical region extraction using an optimal leaf ordering algorithm, and achieved reasonable results. Another approach to segmenting cervical cells is to combine clustering with unsupervised and supervised techniques, such as K-means and SVM. It is achieved through morphological reconstruction and clustering [22]. In parallel, deep learning approaches have gained traction in the division of segmentation, extending the application to cell images. Research by [23] embedded attention modules by learning unique visual patterns through stacked predictive sparse decomposition. They achieved accurate nuclear segmentation through deep learning concepts.

The research by [24] aimed to analyze cervical cell images using a two-level approach, employing the fuzzy C-means (FCM) clustering technique, followed by an artificial neural network on multiclass datasets. In the past decade, diffusion models have gained traction in medical image segmentation. They have achieved remarkable progress, and they work in two ways: Forward and backward diffusion processes. The forward process gradually corrupts the original data x0 (e.g., ground truth segmentation mask) by progressively adding Gaussian noise over T timesteps, transforming it into pure noise. At each timestep t, the noisy data x_t is obtained from x_{t-1} through:

$$q(x_t|x_t) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I \quad (1)$$

whereβ_t is a predefined variance schedule controlling the noise magnitude. This process systematically degrades the data structure until the original information becomes indistinguishable from random noise. Whereas the reverse process learns to denoise and reconstruct clean data from pure noise progressively x_t through iterative refinement. A neural network $\epsilon\theta$ is trained to predict the noise added at each timestep, enabling the recovery of x_{t-1} from x_t.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \quad (2)$$

For instance, a study by Julia et al. [19] developed a novel semantic segmentation method based on diffusion models. A stochastic sampling process was used to generate a distribution of segmentation mask. Another work by Xing et al. [25] built a Diffusion Embedded Network for Volumetric Segmentation, Diff-UNet, which integrates the diffusion model into a standard U-shaped architecture to extract semantic information. To combine the outputs of the diffusion models at each step to combine the outputs of the diffusion models at each step.

More recently, the authors [12] developed MedSegdiff, a Medical image segmentation model based on a diffusion probabilistic model. To better address the challenge of distinguishing lesions or organs from complex backgrounds in medical images, MedSegDiff establishes state-adaptive conditions for each sampling step through dynamic conditional encoding, thereby improving upon the fixed conditioning strategies employed in traditional DDPMs. [13] designed BerDiff, a Bernoulli noise-based diffusion model specifically designed for medical image segmentation. BerDiff leverages the stochastic nature of diffusion models by performing multiple sampling iterations to generate diverse segmentation outputs, enabling explicit visualization of regions of interest and providing richer information.

To enhance the model's ability to discriminate ambiguous regions and perceive target shapes, a growing number of studies have focused on explicitly modeling boundary features within network architectures through dedicated edge-aware modules.

In [26], a framework was introduced that integrates gradient data with classifier probability scores to construct robust Edge-Stop Functions (ESF) for edge-based active contour models. This combined approach overcomes the limitations of traditional gradient-only ESFs, which perform poorly on medical images with ambiguous boundaries, enabling more effective segmentation of blurred boundary regions.

Another study by [27] proposed the Edge Attention Network, which leverages boundary cues as geometric constraints to distinguish foreground from background and improve identification of blurred edges in heterogeneous regions. The framework integrates an Edge Attention Preservation (EAP) module with a Multi-level Pairwise Regression (MPR) module to collaboratively refine target boundaries and enhance segmentation accuracy.

While these approaches validate the importance of explicit boundary modelling for segmentation accuracy, they predominantly rely on fixed gradient operators (Canny, Sobel) or hand-crafted edge features that cannot adapt to the morphological heterogeneity present in cervical cytology images. Cervical nucleus segmentation presents unique challenges: nuclear boundaries frequently exhibit severe ambiguity due to cellular overlap, inconsistent staining quality, and weak contrast between nuclear and cytoplasmic regions. Moreover, nuclei display substantial morphological diversity across different dysplasia grades, ranging from regular circular shapes in normal cells to irregular, bean-shaped, or lobulated contours in high-grade lesions. Traditional boundary detection methods with fixed parameters fail to accommodate this variation. Therefore, our work introduces a fully learnable boundary enhancement mechanism using WCRB, which adaptively weights multi-directional edge components and dynamically fuses boundary information with semantic features via gated mechanisms, yielding task-specific boundary representations optimized for cervical nucleus morphology. Recent 2024-2025 studies on diffusion-based and topology-aware segmentation further motivate the need for adaptive boundary and topology supervision.

3. Method

For clarity and reproducibility, each architectural component is motivated, mathematically formulated, and supported with implementation-level details. Fig. 1 illustrates the complete triplet-branch framework, showing the diffusion backbone, semantic branch, and wavelet-enhanced contour refinement branch, along with their bidirectional information flow. It primarily consists of three synergistic components: the Diffusion Model Backbone, the Semantic Condition Branch, and the Wavelet-Enhanced Contour Refinement Branch.

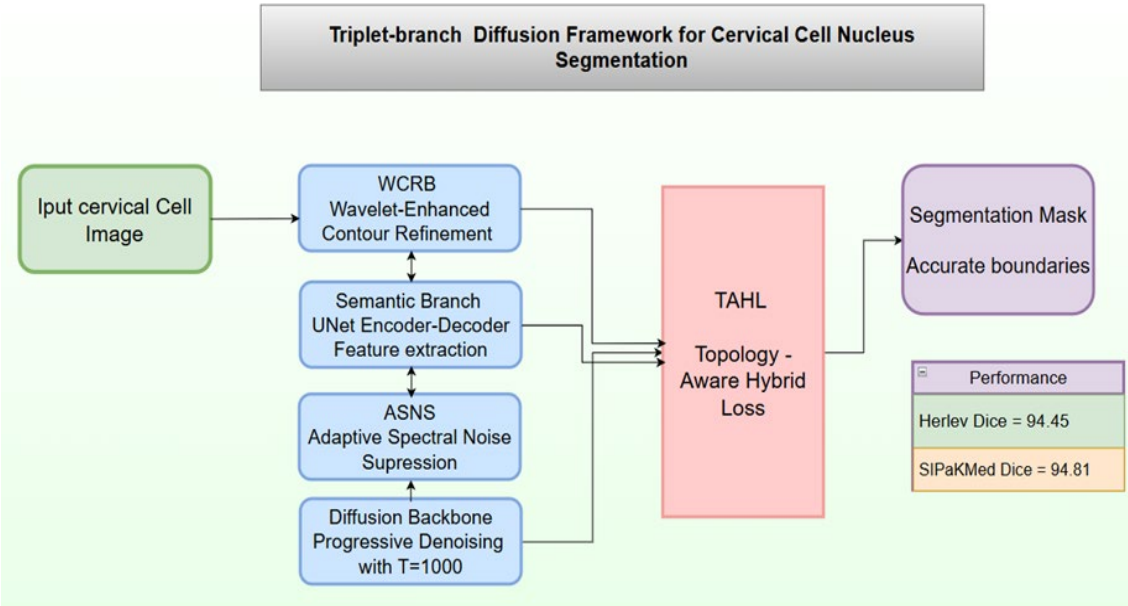


Fig. 1. Proposed Architecture

The Diffusion Model Backbone serves as the core generative engine [28], [29], built on a U-Net-like architecture that progressively reconstructs clean segmentation masks from noisy inputs through iterative denoising [30]. Fig. 2 demonstrates the diffusion backbone architecture.

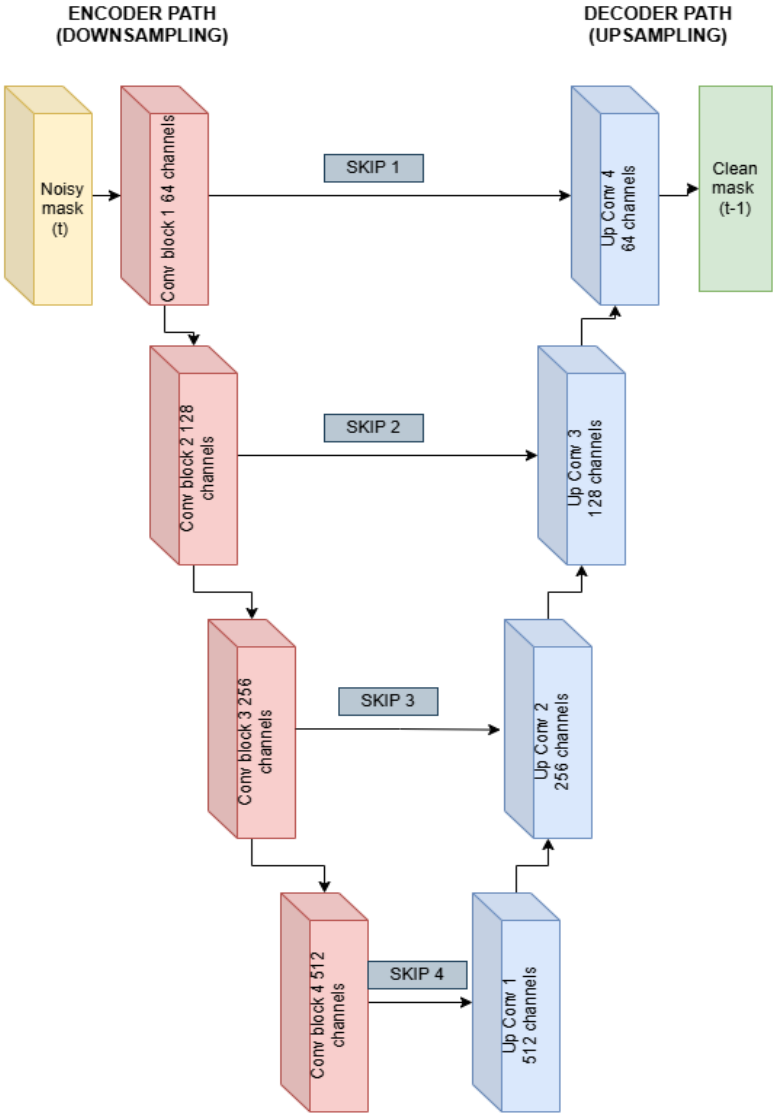


Fig. 2. UNet diffusion backbone

Unlike conventional discriminative segmentation models that directly map images to binary masks, this diffusion-based approach models the joint distribution between input images and target masks, enabling more robust handling of ambiguous regions and morphological variations. The backbone receives concatenated noisy masks and input images as input, simulating the reverse diffusion process through T=1000 timesteps. To mitigate the adverse effects of noise accumulation during this progressive reconstruction, we integrate the ASNS module at multiple encoder levels. This module performs frequency-domain filtering with uncertainty-guided fusion, effectively separating signal from noise while preserving critical structural information. Furthermore, an anchor-based feedback mechanism extracts refined structural features from the denoised representations and feeds them back to the Semantic Condition Branch, establishing bidirectional information coupling that enhances cross-branch guidance and feature discriminability.

To provide rich semantic context and auxiliary supervision, we designed a parallel Semantic Condition Branch based on the U-Net architecture [31]. This branch directly extracts multi-scale semantic features from input cervical cell images without injecting noise, thereby capturing anatomical priors and global contextual information. High-level semantic features from the encoder's deepest layer are injected into

the bottleneck and early decoding stages of the diffusion backbone via cross-layer skip connections, thereby guiding the denoising process with clean, semantically meaningful representations. Additionally, this branch generates an independent segmentation output that serves as an auxiliary supervision signal during training, enabling the model to learn complementary representations through multi-task optimization. The semantic features act as a stabilizing force, anchoring the reconstruction process of the diffusion backbone and preventing it from diverging into semantically implausible solutions.

Furthermore, to explicitly address the challenge posed by blurred and ambiguous nuclear boundaries, a pervasive issue in cervical cytology images, we introduce the Wavelet-Enhanced Contour Refinement Branch (WCRB). Unlike conventional edge detection methods that rely on hand-crafted operators such as Canny or Sobel, WCRB employs a learnable wavelet decomposition approach to capture multi-scale boundary information across different frequency sub-bands. The branch performs 2D Discrete Wavelet Transform (DWT) on input images to decompose them into approximation and detail coefficients (LL, LH, HL, HH), with the detail coefficients encoding horizontal, vertical, and diagonal edge information. The AWCA mechanism learns to dynamically weight these frequency components, generating a Contour Saliency Map (CSM) that highlights boundary regions. Through MCFI modules equipped with gated fusion mechanisms, the boundary information is progressively injected into the decoding stages of both the Diffusion Model Backbone and the Semantic Condition Branch. This multi-path boundary enhancement strategy ensures that contour information is preserved and refined throughout the reconstruction process, significantly improving the model's ability to delineate fine-grained nuclear boundaries even in the presence of cellular overlap and weak staining intensity.

The three branches operate synergistically through carefully designed information flow pathways. The ASNS module denoises the diffusion backbone's features in the frequency domain and provides anchor feedback to the semantic branch, creating a bidirectional coupling that enhances semantic guidance. The WCRB extracts learnable boundary features and fuses them into both the diffusion backbone and semantic branch decoders through gated attention mechanisms, ensuring that edge information is consistently reinforced across all decoding scales. The semantic branch provides clean, high-level contextual features to guide the diffusion backbone's reconstruction while simultaneously producing auxiliary segmentation outputs for multi-task learning. This collaborative architecture enables the model to simultaneously optimize for semantic accuracy, boundary precision, and noise robustness.

During training, the proposed tool is jointly optimized using the proposed TAHL, which extends beyond conventional pixel-wise or region-based loss functions by incorporating explicit topological constraints. TAHL comprises five complementary components: (1) a focal Tversky loss for handling class imbalance and emphasizing false negatives, (2) a persistent homology-based topology loss that penalizes incorrect numbers of connected components and holes, (3) a directional boundary loss that enforces correct orientation of edge gradients, (4) a skeleton completeness loss that ensures contour connectivity, and (5) the standard MSE loss for diffusion noise prediction. By simultaneously optimizing across semantic regions, boundary structures, topological correctness, and diffusion reconstruction, TAHL provides comprehensive supervision that guides the model to generate anatomically plausible segmentation masks with accurate boundaries and preserved topological properties.

The proposed triplet-branch tool achieves multi-scale, multi-domain collaborative optimization through three key innovations: frequency-domain noise suppression with anchor feedback, WCRB, and TAHL. These components work in concert to address the fundamental challenges in cervical nucleus segmentation, including noise interference, boundary ambiguity, and topological inconsistency, while maintaining high computational efficiency and clinical interpretability. The following subsections provide detailed technical descriptions of each novel component.

3.1. Wavelet-Enhanced Contour Refinement Branch

The architecture of WCRB is illustrated in Fig. 3, and its operation encompasses three primary stages: wavelet-based feature extraction with adaptive attention, contour saliency generation with multi-level integration, and residual boundary enhancement.

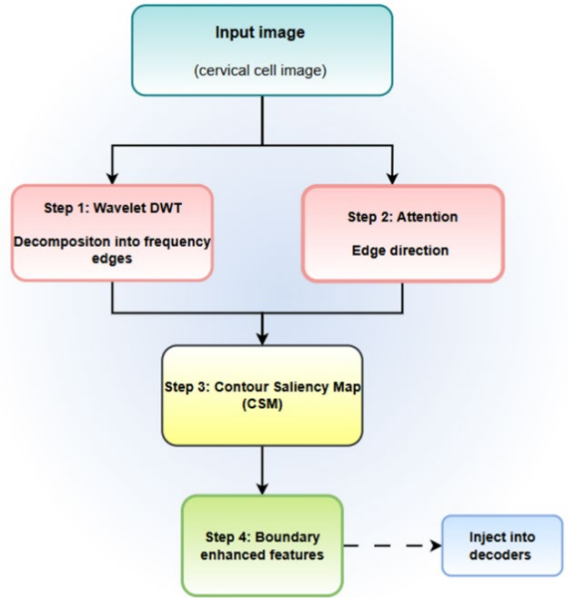


Fig. 3. Wavelet Enhanced Contour Refinement Branch

The first stage applies the 2D Discrete Wavelet Transform (DWT) [32] to decompose the input cervical cell image $I \in \mathbb{R}^{H\times W\times 3}$ into four frequency sub-bands that encode both approximation and detail information at different orientations. The DWT [33] operates by convolving the input image with a pair of complementary filters: a low-pass filter (associated with the scaling function) and a high-pass filter (associated with the wavelet function), followed by down-sampling operations. In this work, we employ the Haar wavelet basis for its computational efficiency and its effectiveness in capturing sharp discontinuities, which are characteristic of cellular boundaries. The decomposition process can be formally expressed as:

$$\mathrm{WT}\ (I) = \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\} \tag{3}$$

where LL denotes the approximation coefficients obtained by applying the low-pass filter in both horizontal and vertical directions, capturing low-frequency content and representing a coarse-scale representation of the input image. LH denotes the horizontal detail coefficients obtained by applying the low-pass filter horizontally and the high-pass filter vertically, primarily encoding vertical edge information. HL denotes the vertical detail coefficients obtained by applying a high-pass filter horizontally and a low-pass filter vertically, thereby capturing horizontal edge structures. HH denotes the diagonal-detail coefficients obtained by applying a high-pass filter in both directions, encoding diagonal edges and corner features.

While wavelet decomposition extracts multi-directional edge information, not all frequency components contribute equally to nuclear boundary detection. Depending on the local cellular morphology and orientation, certain directional edges may be more salient than others. To enable the model to emphasize the most informative frequency components adaptively, we introduce the AWCA mechanism, which learns to weight the detail coefficients according to their relevance to boundary detection. The AWCA mechanism operates in two stages. First, we compute a weighted combination of the absolute magnitude of the three detail coefficients to form an aggregated detail map:

$$W_{\text{detail}} = \alpha \cdot |\mathrm{LH}| + \beta \cdot |\mathrm{HL}| + \gamma \cdot |\mathrm{HH}| \tag{4}$$

where α, β, γ are learnable scalar parameters initialized to $[0.33, 0.33, 0.34]$ to ensure balanced weighting at the start of training. The absolute value operation ensures that edge responses are represented as non-negative magnitudes, consistent with the physical interpretation of edge strength. These learnable coefficients allow the network to automatically determine the relative importance of horizontal, vertical,

and diagonal edges for the specific task of nuclear boundary detection, adapting to the predominant orientations observed in the training data.

Second, to further refine the attention mechanism and enable spatial adaptivity, we learn a channel-wise attention map by applying a 1×1 convolution followed by a sigmoid activation to the concatenated detail coefficients:

$$A_{\text{wavelet}} = \sigma(\text{Conv}_{1\times1}(\text{Concat}(\text{LH}, \text{HL}, \text{HH}))) \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, which constrains the attention values to the range (0, 1), and $Concat(\cdot)$ represents channel-wise concatenation. The 1×1 convolution performs a learned linear combination across channels, enabling the network to identify complex interactions between different directional components. This spatially varying attention map assigns higher weights to frequency regions that exhibit strong boundary characteristics while suppressing regions dominated by noise or texture variations. The combination of learnable coefficient weighting and spatial attention ensures that AWCA can adapt to diverse boundary patterns, including weak edges in low-contrast regions and strong edges at cell-cell interfaces, thereby enhancing the robustness and discriminative power of the extracted boundary features.

Having obtained the adaptively weighted detail map $Wdetail$ and the spatial attention map A_{wavelet}, We proceed to generate a Contour Saliency Map that explicitly highlights boundary regions while suppressing irrelevant background areas. The CSM is computed by fusing the weighted detail information with the spatial attention through element-wise multiplication, followed by a 3×3 convolution to capture local spatial context, and a hyperbolic tangent activation to normalize the output:

$$CSM = tanh(Conv3 \times 3(Wdetail \odot Awavelet)) \quad (6)$$

where $\odot$ denotes element-wise multiplication (Hadamard product). The element-wise multiplication operation serves as a feature-gating mechanism, allowing the spatial attention map to modulate the contribution of each spatial location in the detail map. Regions with high attention values (corresponding to strong boundary evidence) are amplified, while regions with low attention values (corresponding to homogeneous areas or noise) are attenuated. The subsequent 3×3 convolution aggregates information from local neighbourhoods, enabling the network to capture spatially coherent boundary structures rather than isolated edge pixels. This is particularly important in cervical cell images, where nuclear boundaries often form closed contours spanning multiple pixels. The hyperbolic tangent activation function maps the convolution output to the range $(-1, 1)$, providing a normalized representation that facilitates stable gradient flow during backpropagation and prevents feature saturation. The resulting CSM serves as a high-fidelity representation of boundary saliency, encoding both the location and strength of nuclear contours. Unlike binary edge maps produced by classical operators, the CSM retains continuous edge-confidence scores, enabling more nuanced boundary modelling and smooth integration with downstream feature-fusion modules.

Given a feature map $F_{image} \in \mathbb{R}^{H\times W\times C}$ from a decoder layer and the corresponding CSM (appropriately resized to match the spatial dimensions), the MCFI module first computes a gating map G that determines the relative contribution of each information source:

$$G = \sigma\left(\text{Conv}_{3\times3}([F_{\text{image}}, \text{CSM}])\right) \quad (7)$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation, and the 3×3 convolution learns a non-linear mapping from the concatenated features to a gating signal. The sigmoid activation ensures that the gating values lie in the range (0, 1), with values close to 1 indicating that the original feature map should dominate, and values close to 0 indicating that the boundary information should dominate. This adaptive gating mechanism allows the network to selectively emphasize boundary features in regions where edges are prominent (e.g., at nuclear contours) while preserving semantic information in homogeneous regions

(e.g., within nuclear interiors or background areas). The refined feature map F_{refined} is then obtained by blending the gated original features with the boundary-enhanced features:

$$F_{\text{refined}} = G \odot F_{\text{image}} + (1 - G) \odot \text{Conv}_{3\times3}(\text{CSM}) \quad (8)$$

This formulation ensures that the output is a smooth interpolation between the two information sources, controlled by the learned gating map. When G approaches 1, the refined features are dominated by the original semantic representation, preserving the model's ability to capture regional context and texture. Conversely, when G approaches 0, the refined features are driven primarily by the boundary saliency map, sharpening contours, and enhancing edge localization. The 3×3 convolution applied to the CSM before fusion further enriches the boundary representation by capturing local spatial dependencies. The gated fusion mechanism provides a principled and differentiable approach to combining multi-modal information, enabling the network to learn task-optimal fusion strategies during end-to-end training, contrasting with fixed fusion rules that lack the flexibility to adapt to varying boundary characteristics across different images and regions.

To further strengthen the boundary representations and ensure that edge information is consistently propagated through the network, we introduce a residual connection that directly injects boundary features into the refined feature map. Inspired by the success of residual learning in deep neural networks, this design facilitates gradient flow and prevents the degradation of boundary information during feature transformation. The final output of the WCRB is computed as:

$$F_{\text{output}} = F_{\text{refined}} + \lambda_{\text{residual}} \cdot \left(\text{Conv}_{1\times1}(\text{CSM}) \odot F_{\text{image}}\right) \quad (9)$$

where $\lambda_{\text{residual}}$ is a learnable scalar parameter initialized to 0.1, which controls the magnitude of the residual boundary enhancement. The 1×1 convolution projects the CSM into the same channel space as F_{image}, enabling element-wise multiplication to perform channel-wise gating. This residual term acts as an explicit boundary attention mechanism, amplifying feature responses in regions corresponding to nuclear contours while leaving non-boundary regions relatively unaffected.

The learnable scaling parameter $\lambda_{\text{residual}}$ allows the network to automatically balance the strength of the residual connection during training. In early training stages, when boundary features may be noisy or unreliable, the network can downweight the residual term by reducing $\lambda_{\text{residual}}$. As training progresses and boundary predictions become more accurate, $\lambda_{\text{residual}}$ can increase to provide stronger boundary reinforcement. This adaptive behavior enhances training stability and convergence.

The output features F_{output} from the WCRB are injected into the decoder stages of both the Diffusion Model Backbone and the Semantic Condition Branch at three different spatial resolutions: 32×32, 64×64, and 128×128 pixels. This multi-level injection strategy ensures that boundary information is incorporated at multiple scales, allowing the model to refine boundaries progressively from coarse to fine resolutions. At lower resolutions (32×32), the boundary features guide the overall shape and topology of nuclear regions, while at higher resolutions (128×128), they refine pixel-level edge localization and contour smoothness.

The WCRB is designed to be lightweight and computationally efficient, with a total of approximately 0.8 million trainable parameters, which is significantly fewer than typical boundary encoder networks. The 2D DWT is implemented using PyWavelets library with Haar wavelet basis, and the decomposition is performed on grayscale versions of the input images obtained by averaging the RGB channels. The wavelet coefficients are normalized to zero mean and unit variance to ensure stable training dynamics. The AWCA module employs 1×1 convolutions with 64 output channels, followed by batch normalization and sigmoid activation. The gating convolutions in the MCFI module use 3×3 kernels with padding to preserve spatial dimensions, and are also followed by batch normalization layers. All convolutions use He initialization for weight initialization, and biases are initialized to zero.

3.2. Adaptive Spectral Noise Suppression Module

Diffusion models, by their fundamental design, operate through a progressive denoising process that iteratively refines noisy inputs to recover clean target structures. While this iterative refinement mechanism endows diffusion models with remarkable generative capabilities, it also introduces a critical challenge: the accumulation of noise artifacts during the forward diffusion process can corrupt the intermediate feature representations within the encoder pathway, leading to degraded semantic understanding and impaired boundary localization. Unlike conventional discriminative segmentation networks that process clean input images directly, the Diffusion Model Backbone in our framework receives concatenated noisy masks and images as input, where the noise level varies across diffusion timesteps according to a predefined schedule (from t=0 to t=1000). This noise injection, while essential for the generative formulation, introduces high-frequency perturbations that contaminate the feature maps extracted by the encoder, potentially causing the model to hallucinate spurious structures, misidentify background regions as nuclei, or fail to detect genuine boundaries obscured by noise-induced texture variations in densely overlapping cervical cell clusters.

The key innovation of ASNS lies in its dual-domain processing strategy: it simultaneously extracts features in both the spatial domain and the frequency domain then fuses them using an uncertainty-guided weighting mechanism that assigns higher confidence to spatial features in low-noise regions and higher confidence to frequency-filtered features in high-noise regions. This adaptive fusion strategy enables ASNS to preserve fine-grained structural details in clean regions while aggressively suppressing noise in corrupted regions, thereby achieving superior denoising performance compared to fixed-rule filtering approaches.

The ASNS module receives encoder feature $x \in \mathbb{R}^{H\times W\times C}$ from the Diffusion Model Backbone and processes them through two parallel pathways: a spatial stream and a spectral stream. The spatial pathway applies standard convolutional processing with Gaussian Error Linear Unit (GELU) activation and batch normalization to extract local spatial features:

$$F_{spatial} = GELU(BN(Conv_{3\times3}(\mathrm{x}))) \tag{10}$$

where BN(·) denotes batch normalization. This pathway preserves local texture patterns and spatial correlations that are essential for distinguishing cellular structures. The frequency pathway transforms the feature map into the spectral domain using the 2D Discrete Cosine Transform (DCT):

$$F_{freq} = DCT_{2D}(x) \tag{11}$$

The DCT is preferred over the Fourier Transform due to its superior energy compaction for natural images; it concentrates signal energy in fewer low-frequency coefficients, making signal-noise separation more effective. The DCT representation $F_{freq} \in \mathbb{R}^{H\times W\times C}$ is purely real-valued, avoiding the computational overhead of complex arithmetic required by FFT.

To enable adaptive frequency filtering, we apply a channel attention mechanism to the spectral representation. First, the magnitude of the frequency coefficients is couted as $M_{\text{freq}} = |F_{freq}|$. A squeeze-and-excitation style channel attention is then applied to learn importance weights for different frequency channels:

$$W_{\text{channel}} = \sigma\left(\mathrm{FC}_2\left(\mathrm{GELU}\left(\mathrm{FC}_1\left(\mathrm{GAP}(M_{\text{freq}})\right)\right)\right)\right) \tag{12}$$

where GAP(·) denotes global average pooling, FC_1 and FC_2 are fully connected layers with a channel reduction ratio of 16, and $\sigma(\cdot)$ is the sigmoid activation. This produces channel-wise importance weights $W_{channel} \in \mathbb{R}^{1\times1\times C}$.

Unlike fixed frequency masking, ASNS learns spatially-adaptive filters by combining magnitude and phase information. The phase component is computed as the sign of DCT coefficients, denoted as $\angle F_{\text{freq}}$. The adaptive spatial mask is obtained through:

$$M_{\text{adaptive}} = \text{Conv}_{1\times1}\left(\text{Concat}\left(|F_{\text{freq}}|, \angle F_{\text{freq}}\right)\right) \quad (13)$$

The filtered frequency representation is obtained by applying both channel and spatial modulation:

$$F'_{\text{freq}} = F_{\text{freq}} \odot M_{\text{adaptive}} \odot W_{\text{channel}} \quad (14)$$

where $\odot$ denotes element-wise multiplication, with broadcasting applied for W_{channel}. This adaptive filtering mechanism allows the network to selectively suppress noise-dominated frequency components while preserving signal-rich coefficients, with the spatial adaptivity enabling different filtering strategies across different image regions. Subsequently, Uncertainty-Guided Fusion and Anchor Feedback is discussed below.

To adaptively weight the contributions of spatial and frequency features based on local noise characteristics, we estimate spatial uncertainty using local variance. For each spatial location, the local variance is computed within a 5×5 sliding window:

$$U(i,j) = \sqrt{\mathbb{E}_{\Omega(i,j)}[(x-\mu_{\text{local}})^2]} \quad (15)$$

where Ω(i,j) denotes the local neighborhood around pixel (i,j) and μ_{local} are the local mean. The uncertainty map is normalized and passed through a sigmoid modulation to produce spatial weighting:

$$W_{\text{uncertainty}} = \frac{1}{1+\exp(-k\cdot(U-\tau))} \quad (16)$$

where $k = 10$ controls the steepness of the transition, and τ is an adaptive threshold set to the median value of U. This weighting assigns values near 1 to high-uncertainty regions (which require stronger frequency filtering) and near 0 to low-uncertainty regions (which preserve spatial details). The final denoised representation is obtained through uncertainty-guided fusion:

$$F_{\text{denoised}} = W_{\text{uncertainty}} \odot F_{\text{spatial}} + \left(1 - W_{\text{uncertainty}}\right) \odot \text{IDCT}_{2D}\left(F'_{\text{freq}}\right) \quad (17)$$

where $IDCT_2D(\cdot)$ denotes the inverse discrete cosine transform. This formulation enables spatially varying denoising strength, aggressively filtering high-uncertainty regions while preserving details in low-uncertainty regions.

To establish bidirectional coupling with the Semantic Condition Branch, we extract robust structural features from the denoised representation using morphological analysis. The anchor semantics are computed as the difference between max-pooled and min-pooled features:

$$A_{\text{structure}} = \text{MaxPool}_k(F_{\text{denoised}}) - \text{MinPool}_k(F_{\text{denoised}}) \quad (18)$$

where $k = 3$ is the pooling kernel size. This operation effectively extracts edge-strength maps that highlight structural boundaries while suppressing uniform regions and noise-induced texture. The anchor features are then projected through a 1×1 convolution:

$$F_{\text{feedback}} = \text{Conv}_{1\times1}\left(\text{Concat}(A_{\text{structure}}, F_{\text{denoised}})\right) \quad (19)$$

These feedback features F_{feedback} are routed to the corresponding encoder level of the Semantic Condition Branch via cross-branch skip connections, enriching its semantic representations with noise-purified structural information. This bidirectional information flow creates synergistic collaboration

between the diffusion and semantic pathways: the semantic branch receives cleaner structural cues from the denoised diffusion features, while the diffusion branch benefits from the semantic guidance provided by the condition branch. This mutual enhancement is particularly effective in challenging scenarios where noise and semantic ambiguity coexist.

3.3. Topology-Aware Hybrid Loss Function

The effectiveness of deep learning-based segmentation models is fundamentally determined not only by their architecture but also by the optimization objectives that guide their learning. Conventional loss functions employed in medical image segmentation, such as cross-entropy loss or standard Dice loss, implicitly assume that all pixels contribute equally to segmentation quality, focusing primarily on maximizing pixel-wise classification accuracy or region-based overlap between predictions and ground-truth masks. However, this assumption is critically misaligned with the clinical requirements of cervical nucleus segmentation, where different image regions carry vastly different diagnostic significance. Nuclear boundaries, which encode morphological features such as irregular contours, membrane thickening, and nuclear pleomorphism, which are the key indicators of malignant transformation, are far more diagnostically relevant than the homogeneous interior regions of nuclei. A segmentation that achieves 95% pixel-wise accuracy but exhibits systematic boundary erosion or dilation by just 2-3 pixels can lead to significant errors in downstream analyses, including nuclear-to-cytoplasmic ratio estimation, chromatin texture quantification, and automated lesion grading. Furthermore, conventional loss functions fail to enforce topological constraints on the predicted segmentation masks, potentially generating anatomically implausible results such as fragmented nuclei (incorrect number of connected components), spurious holes within nuclear regions (topological defects), or disconnected boundary contours (contour incompleteness).

To provide comprehensive supervision that simultaneously addresses region accuracy, boundary precision, topological correctness, and contour completeness, we propose the TAHL, a multi-component loss function that integrates five complementary objectives, each targeting a specific aspect of segmentation quality. TAHL comprises: (1) a focal Tversky loss that handles severe class imbalance by asymmetrically penalizing false negatives more than false positives, addressing the challenge that nuclei occupy only 10-20% of image area; (2) a persistent homology loss that explicitly preserves topological features by penalizing discrepancies in the number and persistence of connected components and holes between prediction and ground truth; (3) a directional boundary loss that enforces correct orientation of edge gradients, ensuring that predicted boundaries exhibit consistent tangent directions matching the ground truth contours; (4) a skeleton completeness loss that measures the integrity of the morphological skeleton, penalizing boundary discontinuities and fragmentation; and (5) the standard MSE loss for diffusion noise prediction. By jointly optimizing across these five objectives with carefully tuned weighting coefficients, TAHL guides the triplet-branch framework enhanced by WCRB's boundary features and ASNS's denoised representations toward generating segmentation masks that are simultaneously accurate in regional classification, precise in boundary localization, correct in topological structure, and complete in contour connectivity. The TAHL image is presented in Fig. 4.

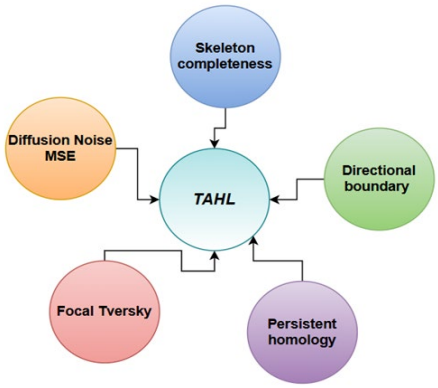


Fig. 4. Topology Aware Hybrid Loss (TAHL)

The first component addresses the severe class imbalance inherent in cervical cell images, where nuclear pixels typically comprise only 10-20% of the total image area. We employ the focal Tversky loss, which asymmetrically penalizes false negatives more heavily than false positives, aligning with clinical priorities in which missing nuclear regions are more detrimental than over-segmentation. The Tversky index is defined as:

$$\mathrm{TI}(P,T) = \frac{TP+\epsilon}{TP+\alpha\cdot FN+\beta\cdot FP+\epsilon} \tag{20}$$

where TP, FN, and FP denote true positives, false negatives, and false positives, respectively, and $\epsilon = 10^{-7}$ ensures numerical stability. We set $\alpha = 0.7$ and $\beta = 0.3$ to prioritize recall over precision. To emphasize hard examples, we apply focal modulation:

$$\mathcal{L}_{\text{tversky}} = (1-\mathrm{TI}(P,T))^{\gamma} \tag{21}$$

with $\gamma = 0.75$ providing moderate focusing without destabilizing training. This focal mechanism down-weights the loss contribution from easy examples (well-classified regions) and focuses the model's attention on difficult regions where predictions are uncertain or incorrect.

The directional boundary loss enforces consistency in edge orientations, ensuring that predicted boundaries not only occupy correct spatial locations but also exhibit proper directional flow. We compute edge responses using Scharr operators, which provide improved rotational symmetry compared to Sobel kernels. The gradient orientations for prediction P and ground truth T are:

$$\theta_{\text{gt}} = \mathrm{atan2}\left(\nabla_y T, \nabla_x T\right), \quad \theta_{\text{pred}} = \mathrm{atan2}\left(\nabla_y P, \nabla_x P\right) \tag{22}$$

The directional boundary loss is restricted to actual boundary regions via a binary mask B, is:

$$\mathcal{L}_{\text{direction}} = \frac{1}{\sum B}\sum_{i,j} B_{i,j}\cdot\left|\theta_{\text{gt}}(i,j)-\theta_{\text{pred}}(i,j)\right| \tag{23}$$

This loss ensures that predicted boundaries exhibit correct orientations, promoting smooth, continuous contours while preventing boundary discontinuities that can arise when edges are predicted with correct positions but incorrect tangent directions.

The second major component explicitly preserves topological correctness using persistent homology, a mathematical framework from algebraic topology that characterizes the birth and death of topological features across different scales. For binary segmentation masks, we compute persistence diagrams PD_{pred} , PD_{gt} that encode the lifespan of connected components (0-dimensional features) and holes (1-dimensional features). Each point in a persistence diagram represents a topological feature, with its coordinates indicating the scale at which the feature appears (birth) and disappears (death). Features with long persistence (large death-birth difference) represent significant structural elements, while short-lived features typically correspond to noise or minor irregularities. The topological loss measures the Wasserstein distance between these diagrams:

$$\mathcal{L}_{\text{topology}} = W_2\left(PD_{\text{pred}}, PD_{\text{gt}}\right) \tag{24}$$

3.4. Differentiable Implementation of Persistent Homology Loss

Although classical persistent homology computation is non-differentiable, we employ a differentiable proxy by backpropagating through the Wasserstein distance between persistence diagrams. Persistence diagrams are computed using the Ripser library, while gradients are propagated using a differentiable Wasserstein loss formulation implemented in PyTorch, enabling stable end-to-end training. This loss penalizes incorrect numbers of connected components (e.g., a single nucleus incorrectly split into two fragments), the presence of spurious holes (e.g., false voids within solid nuclear regions), and merging of adjacent nuclei (e.g., two overlapping nuclei incorrectly segmented as one). The persistence diagram computation is implemented using the Ripser library with cubical complex representation, which

efficiently computes persistent homology for 2D binary images. The Wasserstein-2 distance provides a geometrically meaningful metric for comparing persistence diagrams, ensuring that the optimization not only matches the number of topological features but also their persistence (significance).

The fourth component measures skeleton completeness to ensure contour connectivity and structural integrity. The morphological skeleton represents the medial axis of a shape, which is the locus of centers of maximal inscribed circles, and its completeness indicates boundary integrity and the absence of fragmentation. We extract skeletons using morphological thinning, which iteratively erodes the boundary while preserving topology until only the skeleton remains. The skeleton completeness loss computes the overlap between predicted and ground truth skeletons:

$$\mathcal{L}_{\text{skeleton}} = 1 - \frac{2 \cdot \left| S_{\text{pred}} \cap S_{\text{gt}} \right| + \epsilon}{\left| S_{\text{pred}} \right| + \left| S_{\text{gt}} \right| + \epsilon} \tag{25}$$

where S_{pred} = Skeleton(P) and S_{gt} = Skeleton(T). This loss penalizes fragmented or incomplete boundaries that would result in disconnected skeleton branches, encouraging the model to produce continuous, well-connected contours. Unlike the topological loss that focuses on global structural properties (number of components and holes), the skeleton loss specifically targets local connectivity. It ensures that boundaries form complete, unbroken loops around nuclear regions.

The fifth component is the standard MSE loss for diffusion noise prediction, ensuring accurate denoising in the generative process:

$$\mathcal{L}_{\text{diff}} = \| \epsilon - \hat{\epsilon} \|_2^2 \tag{26}$$

here ε and $\hat{\epsilon}$ denote true and predicted noise, respectively. This loss is essential for maintaining the diffusion model's generative capability, ensuring that the reverse diffusion process accurately removes noise at each timestep.

The complete TAHL integrates these five components with carefully tuned weights:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}} \cdot \mathcal{L}_{\text{diff}} + \lambda_{\text{tversky}} \cdot \mathcal{L}_{\text{tversky}} + \lambda_{\text{topology}} \cdot \mathcal{L}_{\text{topology}} + \lambda_{\text{direction}} \cdot \mathcal{L}_{\text{direction}} + \lambda_{\text{skeleton}} \cdot \mathcal{L}_{\text{skeleton}} \tag{27}$$

The loss weights were selected via grid search over the range {0.1, 0.3, 0.5, 1.0, 2.0} on the validation set. Sensitivity analysis showed that λ_{tversky} primarily affected region accuracy, while topology and skeleton losses mainly influenced boundary integrity. The weighting coefficients are set as λ_{diff} = 1.0, λ_{tversky} = 2.0, $\lambda_{\text{topology}}$ = 0.5, $\lambda_{\text{direction}}$ = 0.3, and $\lambda_{\text{skeleton}}$ = 0.4, determined through systematic grid search on a validation set. The higher weight for Tversky loss reflects its primary role in driving overall segmentation accuracy, while the topology, direction, and skeleton losses serve as regularization terms that enforce geometric and structural constraints.

During training, different branches receive different supervision signals to leverage their complementary strengths. The diffusion backbone is supervised by $\mathcal{L}_{\text{diff}}$ for noise prediction at each timestep, ensuring accurate modeling of the reverse diffusion process. The Semantic Condition Branch's auxiliary output receives supervision from all components except $\mathcal{L}_{\text{diff}}$ as this branch operates on clean images without noise injection. The final diffusion-generated output is supervised by the complete TAHL, ensuring that the ultimate prediction satisfies all geometric, topological, and boundary constraints. This multi-level supervision strategy ensures that both intermediate semantic representations and final outputs are jointly optimized toward clinically meaningful objectives.

The training proceeds in an end-to-end manner, with gradients from all five loss components backpropagating through the triplet-branch architecture. The collaborative optimization enables the model to balance multiple competing objectives: the Tversky loss drives high recall and overall accuracy;

the topology loss prevents structural implausibility; the directional loss ensures smooth, continuous boundaries; the skeleton loss enforces connectivity; and the diffusion loss maintains generative quality. By combining TAHL with WCRB's learnable boundary features and ASNS's noise-robust representations, the triplet-branch AI tool achieves comprehensive optimization across all critical dimensions of segmentation quality: region accuracy through asymmetric Tversky loss, topological correctness through persistent homology, boundary orientation through directional gradients, contour completeness through skeleton matching, and generative fidelity through diffusion noise prediction.

4. Results and Discussion

This section validates the proposed methodology through comprehensive experiments. It presents the dataset details, implementation environment, cross-dataset results, and ablation study, followed by a brief discussion of limitations.

4.1. Datasets

To comprehensively evaluate the effectiveness of the proposed triplet-branch tool, experiments were conducted on two publicly available cervical cell image datasets: Herlev and SIPaKMeD. These datasets represent diverse staining protocols and imaging conditions commonly encountered in clinical cervical cancer screening, thereby enabling robust assessment of the model's generalization capability across different data distributions.

4.1.1. Herlev Dataset

The Herlev dataset is a widely recognized benchmark for cervical cell analysis, comprising 917 single-cell images acquired from Pap smear slides. The images were obtained at the Department of Pathology, Herlev University Hospital, Denmark, using the Papanicolaou (Pap) staining technique. Each image contains a single cervical cell with a clearly visible nucleus and cytoplasm regions. The dataset includes seven cell classes: normal superficial squamous, normal intermediate squamous, normal columnar, mild dysplasia (CIN1), moderate dysplasia (CIN2), severe dysplasia (CIN3), and carcinoma in situ. Original images exhibit non-uniform dimensions ranging from 256×256 to 896×768 pixels. For this study, all images were resized to 128×128 pixels to maintain consistency with the model's input requirements. Each image is accompanied by expert-annotated segmentation masks delineating the nuclear region. The dataset was randomly partitioned into training, validation, and test sets in a 70:15:15 ratio, yielding 642 images for training, 138 for validation, and 137 for testing. Details are presented in Table 1.

Table 1. Dataset summary and statistics

Dataset	Images	Staining	Original Size	Resized	Classes	Train	Val	Test
SIPaKMeD	4049	Pap	768×768	128×128	5	2834	608	607
Herlev	917	Pap	256×256 to 896×768	128×128	7	642	138	137

4.1.2. SIPaKMeD Dataset

The SIPaKMeD (Single Image Per Kokytos Medical) dataset is a large-scale collection specifically designed for cervical cell classification and segmentation tasks. The dataset comprises 4,049 isolated cell images extracted from 966 cluster cell images of Pap smear slides. Images were acquired at the Pathology Department of IASO Hospital, Athens, Greece, using standardized Papanicolaou staining protocols with consistent imaging conditions. The dataset categorizes cells into five classes: superficial-intermediate, parabasal, koilocytotic, dyskeratotic, and metaplastic. Each image has dimensions of 768×768 pixels and includes corresponding ground-truth segmentation masks for the nucleus and cytoplasm regions. For nuclear segmentation, only nuclear masks were used in this study. All images were resized to 128×128 pixels for training and evaluation. Following standard practice, the dataset was split into training, validation, and test sets with a ratio of 70:15:15, yielding 2,834 training images, 608 validation images, and 607 test images.

4.1.3. Data Preprocessing

Standard data preprocessing techniques were applied to both datasets to ensure consistency and improve model robustness. Preprocessing steps included intensity normalization to the [0, 1] range, random horizontal and vertical flips for data augmentation during training, and random rotation within ±15 degrees to account for variations in cell orientation. All ground-truth masks were binarized using a threshold of 0.5, with pixel values set to 1 for nuclear regions and 0 for background. Sample images from the SIPaKMeD dataset are given in Fig. 5.

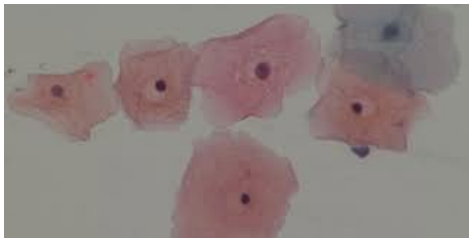


Fig. 5. Sample image from SIPaKMeD dataset

4.2. Implementation

The proposed triplet-branch AI tool was implemented using PyTorch 2.0.1 and trained on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. All experiments were conducted under Ubuntu 22.04 LTS with CUDA 12.1 and cuDNN 8.9. The network was optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-5}. A cosine-annealed learning rate scheduler was employed to gradually reduce the learning rate over the training period, with a minimum rate of 1×10^{-6}. The batch size was set to 8 due to memory constraints, and the model was trained for 200 epochs. Early stopping with a patience of 20 epochs was applied based on validation set performance to prevent overfitting. The diffusion process employed a linear noise schedule with T = 1000 timesteps, with variance parameters β_t increasing linearly from $\beta_1 = 1\times10^{-4}$ to $\beta_t = 0.02$, following the standard DDPM formulation.

The architecture of the Diffusion Model Backbone and Semantic Condition Branch follows a U-Net structure with four encoding and four decoding levels. The initial convolutional layer has 64 channels, which progressively doubles at each down-sampling stage (64 → 128 → 256 → 512), with corresponding spatial resolutions of 128×128, 64×64, 32×32, and 16×16 pixels. Each encoder and decoder block consists of two residual blocks with group normalization and GELU activation. Self-attention mechanisms are integrated into the 16×16-resolution bottleneck to capture long-range dependencies. The WCRB employs a 2-level Haar wavelet decomposition, and the ASNS module is deployed at three encoder levels (128×128, 64×64, 32×32) with channel-wise reduction ratio r = 16 in the spectral attention mechanism. The total number of trainable parameters in the complete framework is approximately 28.7 million.

The loss function weighting coefficients were set as follows: $\lambda_{\text{diff}} = 1.0$, $\lambda_{\text{tversky}} = 2.0$, $\lambda_{\text{topology}}$ y = 0.5, $\lambda_{\text{direction}} = 0.3$, and $\lambda_{\text{skeleton}} = 0.4$. For the focal Tversky loss, the asymmetry parameters were configured as $\alpha = 0.7$ and $\beta = 0.3$, with focal exponent $\gamma = 0.75$. During inference, the diffusion model performs 1000 reverse diffusion steps to generate the final segmentation mask, with an average inference time of approximately 1.8 seconds per image on the RTX 4090 GPU. To quantitatively evaluate segmentation performance, three widely adopted metrics were employed. The Dice Similarity Coefficient (Dice) measures the overlap between the predicted segmentation P and ground truth T, defined as:

$$Dice(P,T) = \frac{2|P \cap T|}{|P|+|T|} \tag{28}$$

where values range from 0 (no overlap) to 1 (perfect agreement). The Intersection over Union (IoU), also known as the Jaccard index, quantifies the ratio of intersection to union:

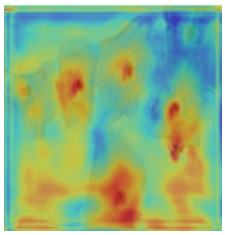
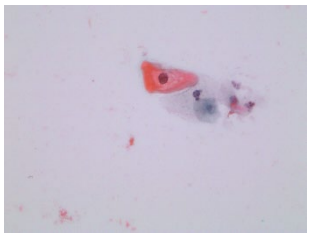
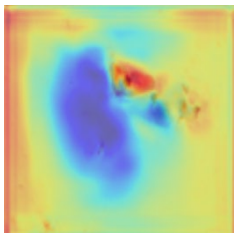
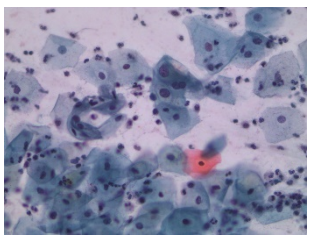
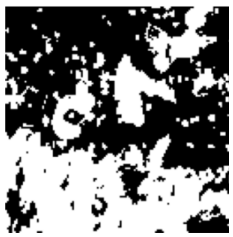
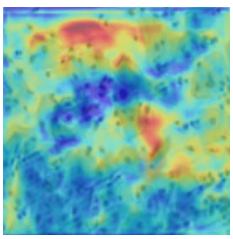
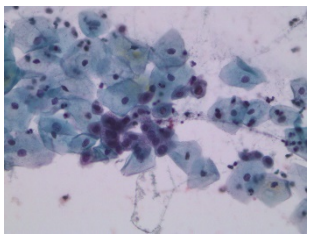
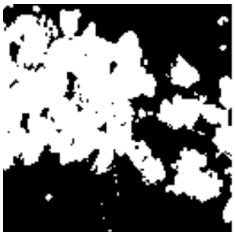
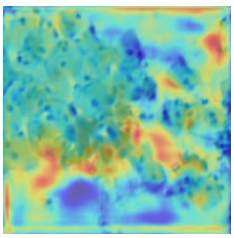
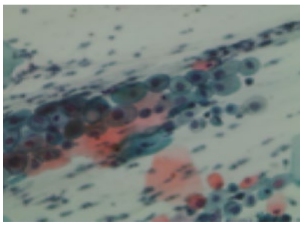
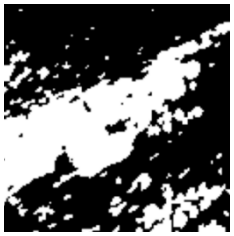
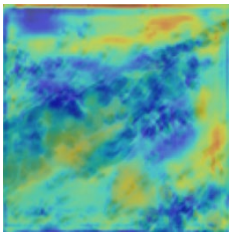
$$IoU(P,T) = \frac{|P \cap T|}{|P \cup T|} \tag{29}$$

The IoU is more sensitive to segmentation errors than Dice and provides a stricter evaluation criterion. The 95th-percentile Hausdorff Distance measures boundary localization accuracy by computing the maximum distance between surface points of the predicted and ground-truth masks. Specifically, HD95 computes the 95th percentile of the distances from each point on the predicted boundary to the nearest point on the ground-truth boundary, thereby excluding outliers while capturing typical boundary errors. Lower HD95 values indicate better boundary precision, which is critical for clinical applications requiring accurate nuclear morphometry. All metrics were computed per image and averaged across the test set to obtain final performance scores.

4.3. Results

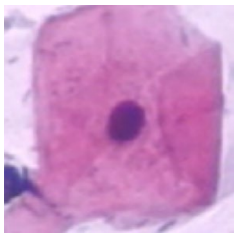
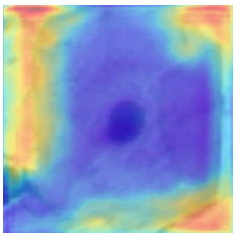
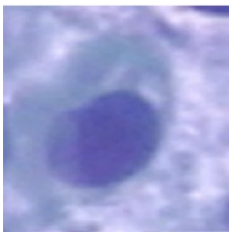
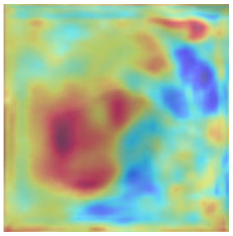
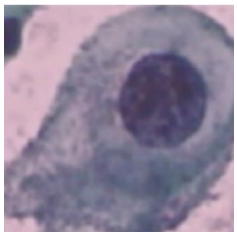
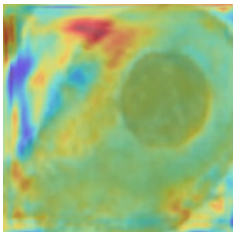
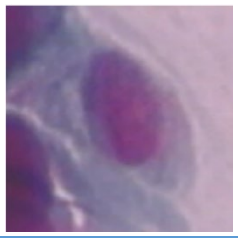
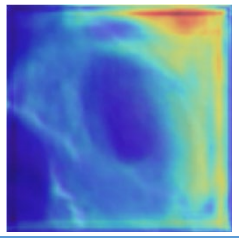
Tables 2 and 3 present qualitative examples of the segmentation results obtained in this study. Table 2 showcases the original images, the generated masks, and the overlay results for several sample images from the SIPaKMeD dataset. These examples illustrate how the segmentation model identifies and highlights relevant cellular regions in the images.

Table 2. Original image, generated masks, and overlay from SIPaKMeD sample images.

Image	Input image	Mask generated	Overlay
SIPaKMeD 1			
SIPaKMeD 2			
SIPaKMeD 3			
SIPaKMeD 4			
SIPaKMeD 5			

Meanwhile, Table 3 displays similar results for sample images from the Herlev dataset, including the input images, the generated masks, and the corresponding overlay visualizations. The overlay images combine the original images with the predicted masks to provide a clearer visual representation of how accurately the model segments the target regions across datasets.

Table 3. Original image, generated masks, and overlay from Herlev sample images.

Image	Input image	Mask generated	Overlay
Herlev Image 1			
Herlev Image 2			
Herlev Image 3			
Herlev Image 4			

Across architectural paradigms, diffusion-based methods generally outperform CNN- and Transformer-based approaches, with the top four performers all employing diffusion models. This superiority stems from the generative formulation's ability to model the joint distribution between images and masks rather than merely learning discriminative decision boundaries.

Statistical significance tests (paired t-test) confirm that the improvements over the best baseline (BerDiff) are statistically significant across all metrics on both datasets ($p < 0.01$), validating that the observed gains are not due to random variation but rather systematic enhancements introduced by the proposed components. The consistent performance across datasets with different characteristics: Herlev contains single-cell images with 917 samples, while SIPaKMeD comprises 4,049 isolated cell images with greater morphological diversity, which further demonstrates the tool's strong generalization capability and robustness to varying data distributions.

4.4. Ablation Studies

To systematically investigate the contribution of each proposed component and validate the design choices within the triplet-branch tool, we conducted comprehensive ablation studies on both the Herlev and SIPaKMeD datasets. The ablation experiments are organized into two categories: 1) Architectural

component ablation, which evaluates the impact of WCRB, ASNS, and their combination; and 2) Loss function ablation, which analyzes the contribution of each term in the TAHL objective.

4.4.1. Architectural Component Ablation

We ran experiments across five architectural variations. Firstly, a baseline model was deployed, after which WCRB and ASNS were added to the baseline individually and together. Finally, the full proposed model WCRB + ASNS + TAHL is experimented with, as shown in Table 4.

Table 4. Results for Architectural ablation, evaluating the impact of WCRB, ASNS, and TAHL.

Configuration	Herlev		SIPaKMeD	
	Dice↑	HD95↓	Dice↑	HD95↓
Baseline (Diffusion + Semantic only)	0.9166	9.6312	0.9203	11.2458
Baseline + WCRB	0.9312	8.1245	0.9345	9.8634
Baseline + ASNS	0.9278	9.0127	0.9314	10.6723
Baseline + WCRB + ASNS	0.9389	7.8923	0.9423	9.3567
Full Model (+ TAHL)	0.9445	7.3124	0.9481	8.9267

The baseline model achieves Dice scores of 91.66% and 92.03% on Herlev and SIPaKMeD, respectively, which are comparable to MedSegDiff (91.66% on Herlev), confirming that our diffusion backbone implementation is sound. Adding WCRB alone yields the most substantial improvement in boundary metrics, reducing HD95 by 1.51 pixels (15.7%) on Herlev and 1.38 pixels (12.3%) on SIPaKMeD, while improving Dice by 1.46 and 1.42 percentage points, respectively. This validates our hypothesis that learnable wavelet-based boundary detection is significantly more effective than traditional fixed edge operators. The pronounced HD95 improvement confirms that WCRB successfully addresses the boundary-ambiguity challenge, which is particularly severe in cervical cell images with overlapping structures and weak staining. Fig. 6 gives a pictorial representation of the architectural ablation outcome.

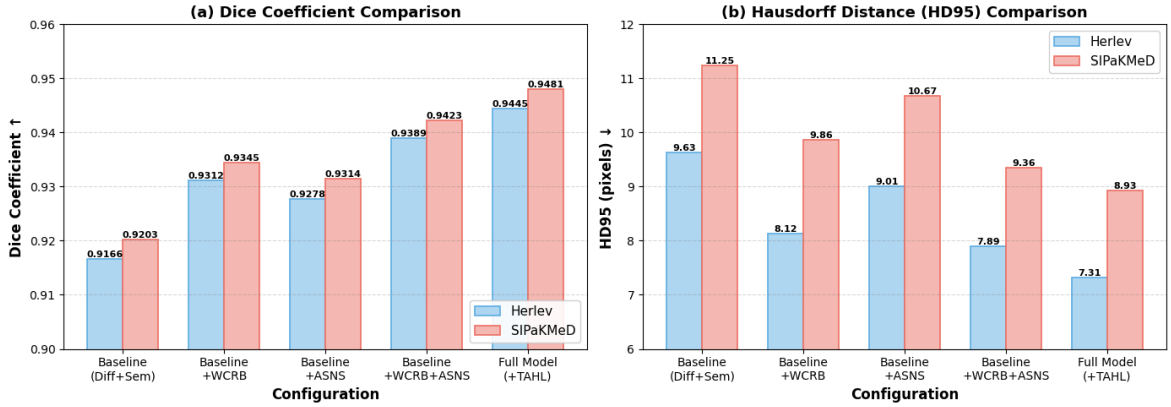


Fig. 6. Result visualization of architectural ablation

Incorporating ASNS alone (without WCRB) improves Dice by 1.12 percentage points at Herlev and 1.11 at SIPaKMeD, with moderate reductions in HD95 of 0.62 and 0.57 pixels, respectively. While ASNS's improvements are less dramatic than WCRB's for boundary metrics, its contribution to overall segmentation accuracy is substantial. The relatively smaller HD95 improvement is expected, as ASNS primarily targets noise suppression in feature representations rather than explicit boundary enhancement. However, the consistent Dice improvements across both datasets demonstrate that cleaner, denoised features enable more accurate region classification and reduce false positives caused by noise-induced texture artifacts.

Fig. 7 shows the radar plot of the Dice coefficient for all models. The combination of WCRB and ASNS (without TAHL) achieves Dice scores of 93.89% and 94.23%, with HD95 values of 7.89 and 9.36 pixels on Herlev and SIPaKMeD, respectively. Notably, the combined improvement (baseline →

WCRB+ASNS: +2.23% Dice on Herlev) slightly exceeds the sum of individual contributions (WCRB alone: +1.46%, ASNS alone: +1.12%), indicating synergistic interaction between the two components. This synergy can be attributed to ASNS providing cleaner features that enable WCRB's boundary attention mechanisms to operate more effectively. In contrast, WCRB's boundary features help ASNS preserve edge information more effectively during spectral filtering.

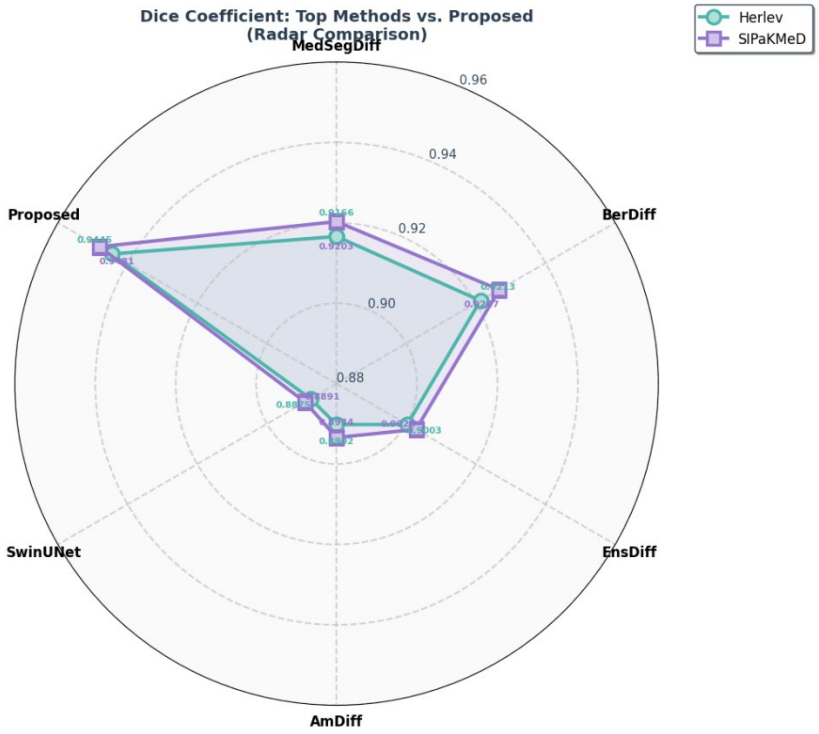
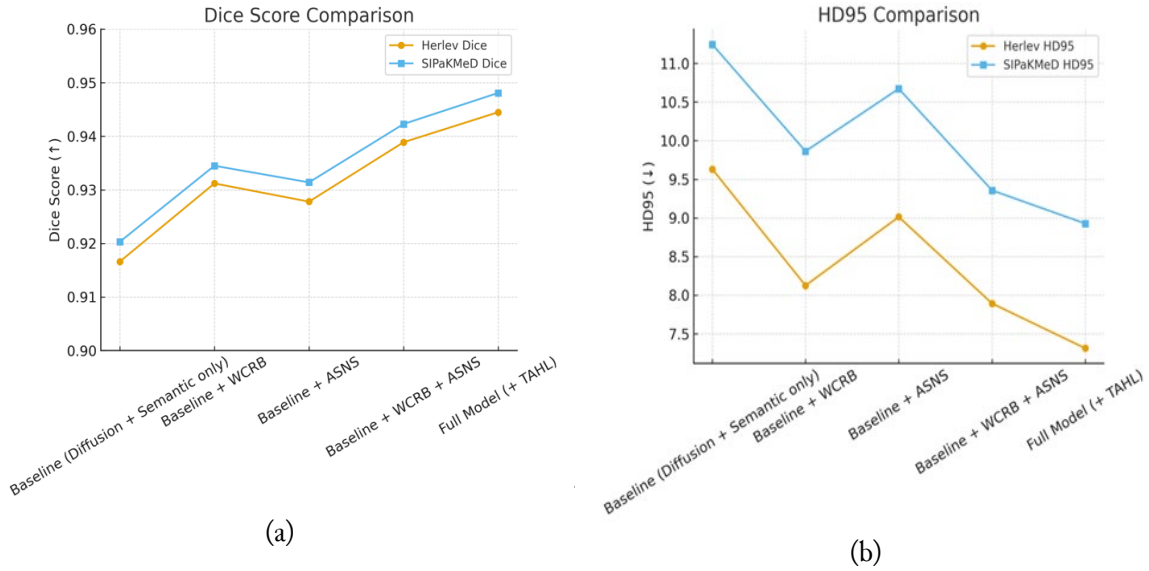


Fig. 7. Radar chart of Dice scores

The full model incorporating TAHL loss yields additional gains of 0.56 and 0.58 percentage points in Dice on Herlev and SIPaKMeD, with further HD95 reductions to 7.31 and 8.93 pixels, respectively. These improvements, while less substantial than the architectural enhancements, are critical to achieving topologically correct segmentation. The infographic is given in Fig. 8.



(a) (b)

Fig. 8. Dice score and HD95 trend in ablation

4.4.2. Ablation study on Loss function

To evaluate the contribution of each component within the proposed TAHL objective, we conducted ablation experiments by systematically adding loss terms to the baseline. The results of loss ablation are presented in Table 5.

Table 5. Results obtained from conducting an ablation on the loss component

Loss Configuration	Dice↑	HD95↓
$\mathcal{L}_{diff}$ only	0.9166	9.6312
$\mathcal{L}_{diff}$ + Dice	0.9267	8.8934
$\mathcal{L}_{diff}$ + Dice + $\mathcal{L}_{hd}$	0.9312	8.2156
$\mathcal{L}_{diff}$ + $\mathcal{L}_{tversky}$	0.9298	8.6745
$\mathcal{L}_{diff}$ + $\mathcal{L}_{tversky}$ + $\mathcal{L}_{topology}$	0.9356	8.1823
$\mathcal{L}_{diff}$ + $\mathcal{L}_{tversky}$ + $\mathcal{L}_{topology}$ + $\mathcal{L}_{direction}$	0.9389	7.8912
$\mathcal{L}_{diff}$ + $\mathcal{L}_{tversky}$ + $\mathcal{L}_{topology}$ + $\mathcal{L}_{direction}$ + $\mathcal{L}_{skeleton}$	0.9423	7.5634
TAHL (All components, full model)	0.9445	7.3124

Starting from diffusion loss alone ($\mathcal{L}_{diff}$), which achieves a 91.66% Dice score, adding a standard Dice loss for the semantic branch improves performance to 92.67%, demonstrating the value of auxiliary supervision. Tversky loss is a generalized loss function for imbalanced segmentation tasks, particularly effective when one class (e.g., nuclei in cervical images) occupies a much smaller area than the background. Replacing the standard Dice with the focal Tversky loss ($\mathcal{L}_{tversky}$) yields a Dice of 92.98%, a 0.31 percentage-point improvement that confirms the effectiveness of asymmetric false-negative penalization for handling class imbalance. Topology loss makes sure the predicted segmentation has the same shape and connectivity as the ground truth. The subsequent addition of a topology loss ($\mathcal{L}_{topology}$) increases the Dice score to 93.56% and reduces the HD95 to 8.18 pixels. This 0.58 percentage point gain validates that explicit topological constraints prevent fragmentation artifacts and improve structural coherence, particularly in challenging cases with clustered or overlapping nuclei. The visual representation is presented in Fig. 9. The trend is seen in Fig. 10.

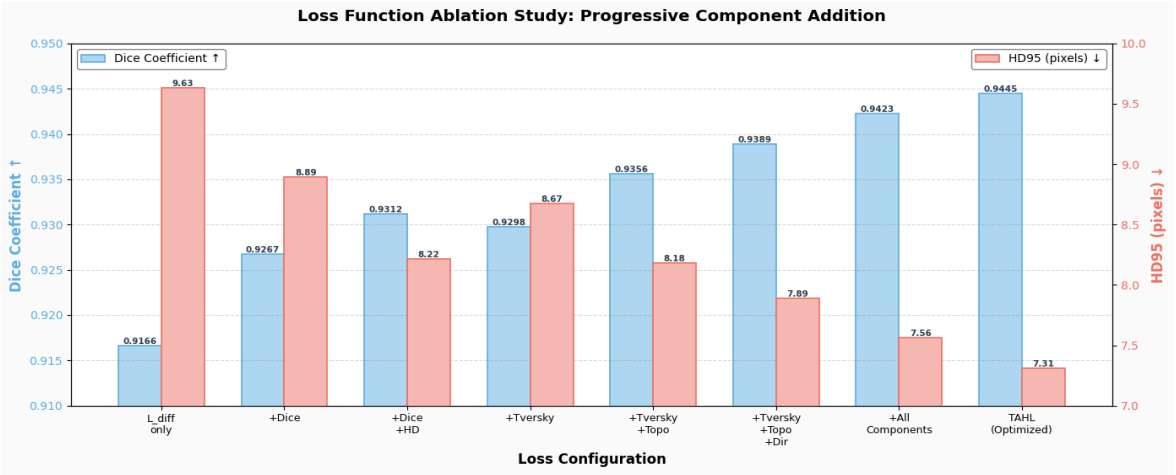


Fig. 9. Result graph of loss function ablation study.

Incorporating directional boundary loss ($\mathcal{L}_{direction}$) further improves Dice to 93.89% and reduces HD95 to 7.89 pixels, with the HD95 reduction being more pronounced (0.29 pixels) than the Dice gain (0.33 points). This pattern is consistent with $\mathcal{L}_{direction}$'s design goal of enforcing correct boundary orientations, which directly impacts boundary localization metrics more than region overlap metrics. Skeleton completeness loss ensures that the predicted object boundaries form continuous, connected contours without missing parts. Finally, adding skeleton completeness loss ($\mathcal{L}_{skeleton}$) yields 94.23% Dice and 7.56 pixels HD95, demonstrating that ensuring contour connectivity provides incremental but meaningful improvements.

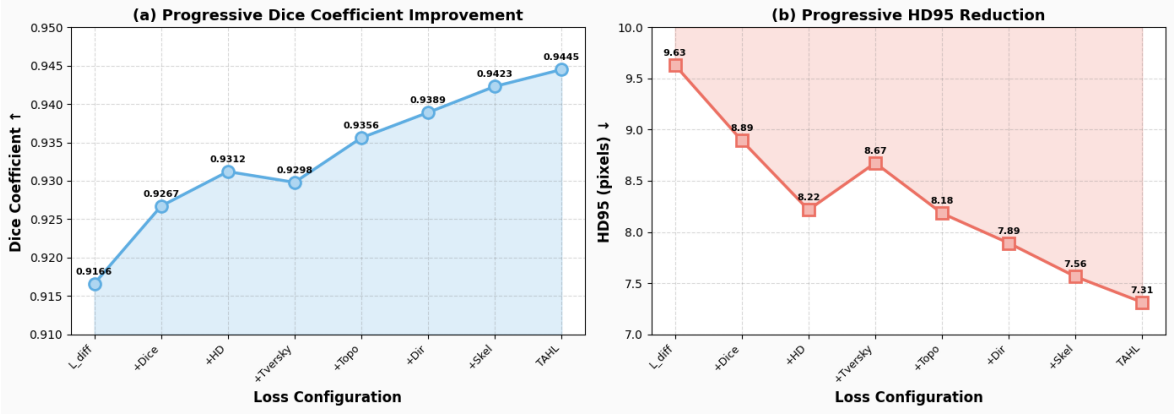


Fig. 10. Dice coefficient and HD 95 trend over loss function ablation study

The full TAHL configuration, which includes all five components with optimized weighting coefficients (λ_{diff} f=1.0, λ_{tversky} =2.0, $\lambda_{\text{topology}}$ =0.5, $\lambda_{\text{direction}}$ =0.3, $\lambda_{\text{skeleton}}$ n=0.4), achieves the best performance of 94.45% Dice and 7.31 HD95. Comparing the full model to the configuration with all five loss terms but equal weights (not shown in the table: Dice=94.18%, HD95=7.68), we observe that the tuned weights provide an additional 0.27 percentage points in Dice and 0.37 pixels in HD95, confirming the importance of proper loss balancing. The systematic performance improvements across the incremental addition of loss terms demonstrate that each component addresses a distinct aspect of segmentation quality, and their combination achieves comprehensive optimization that no single loss can provide.

These ablation studies conclusively demonstrate that the superior performance of the proposed tool stems from the synergistic integration of all three architectural components (WCRB, ASNS, and their incorporation into the diffusion-semantic architecture) combined with the multi-objective TAHL supervision. Removing any component results in measurable performance degradation, validating the necessity of each design choice. The results also provide clear guidance for practitioners: WCRB offers the largest individual gain and should be prioritized when computational resources are limited, while the full configuration delivers optimal performance when all components are deployed jointly.

4.5. Qualitative analysis

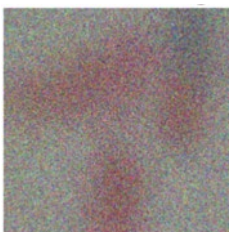
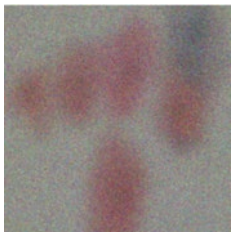
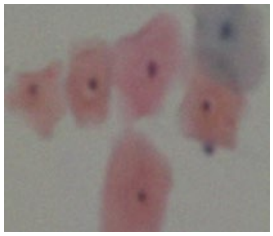
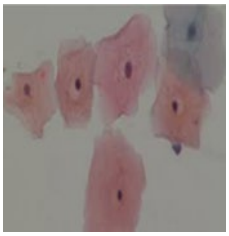
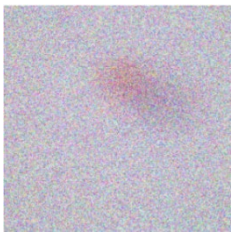
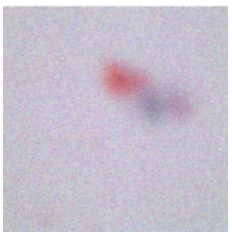
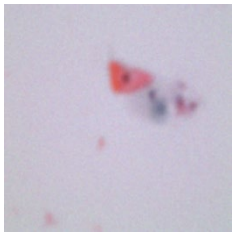
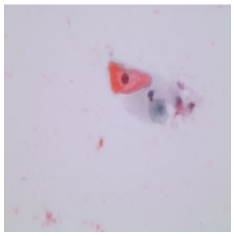
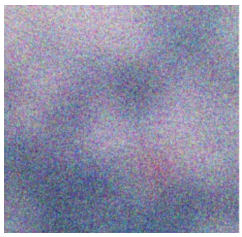
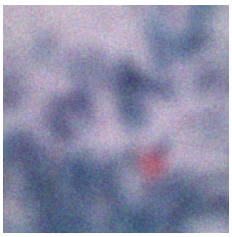
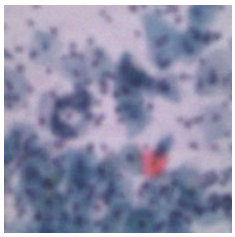
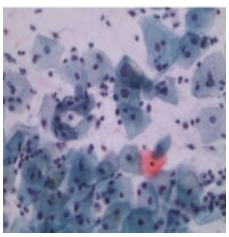
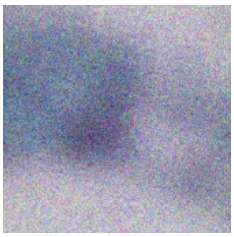
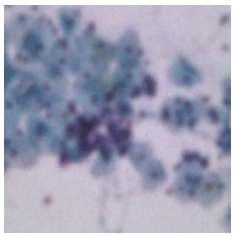
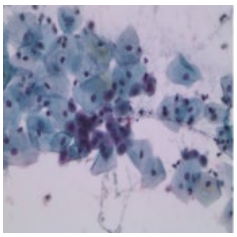
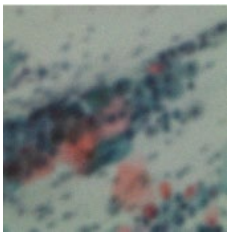
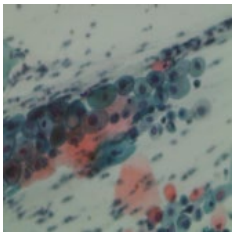
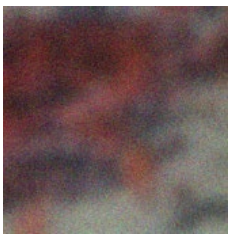
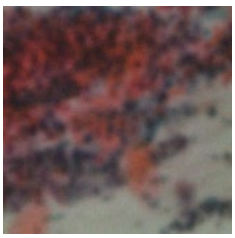
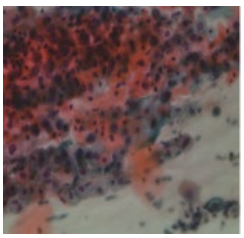
While quantitative metrics provide objective performance assessment, visual inspection of segmentation results offers critical insights into the model's behavior in challenging scenarios. It helps identify failure modes that aggregate statistics may not fully capture. In this section, we present qualitative comparisons with representative baseline methods and visualize the diffusion model's iterative refinement process to demonstrate the progressive boundary enhancement achieved by the proposed framework.

4.5.1. Visualization of Diffusion Process

To provide insight into how the proposed tool progressively refines segmentation masks through the reverse diffusion process, Table 6 visualizes intermediate outputs at five representative timesteps (t = 750, 500, 250, 0) for three test cases. As the reverse diffusion proceeds to t = 750, vague blob-like structures begin to emerge, indicating that the model has started to recover coarse-scale spatial information about nuclear locations. However, boundaries remain extremely blurred, and regional extents are imprecise. By t = 500, nuclear regions become clearly distinguishable from the background, with approximate boundary locations established. However, the contours exhibit significant irregularities, and fine-scale details such as concavities are not yet resolved. This intermediate stage demonstrates the hierarchical refinement strategy inherent in diffusion models, where global structure is recovered before local details. Progressing to t = 250, boundaries become substantially sharper and smoother, with most major morphological features correctly represented. The visible improvement

between t = 500 and t = 250 can be attributed to the increasing influence of WCRB's boundary features and ASNS's denoised representations as noise levels decrease in later diffusion steps.

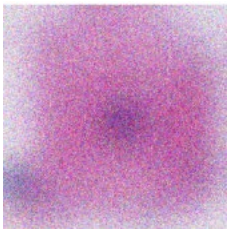
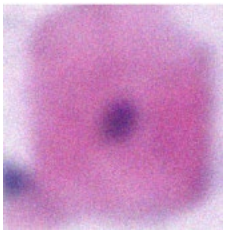
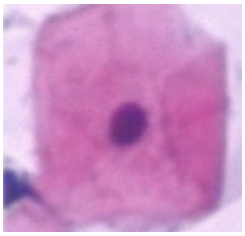
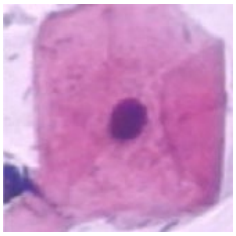
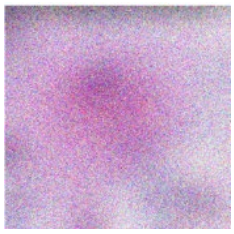
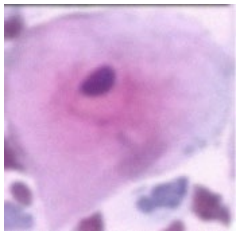
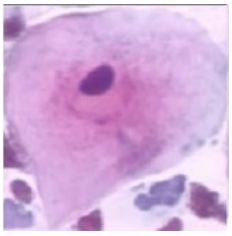
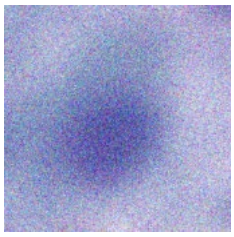
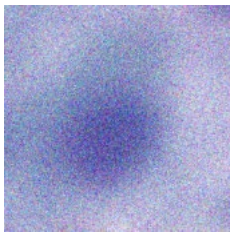
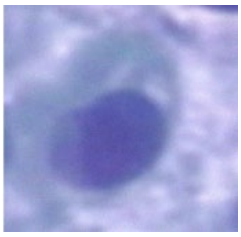
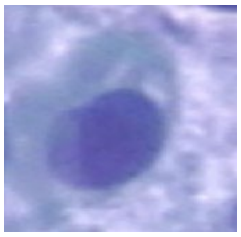
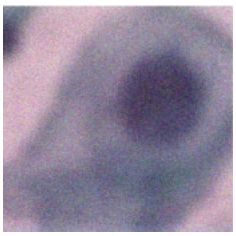
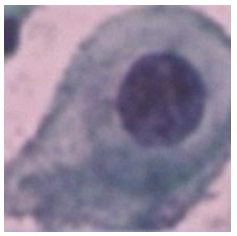
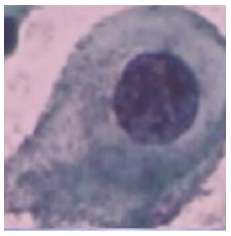
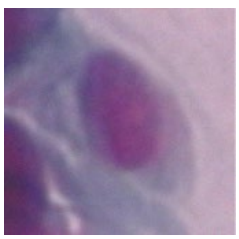
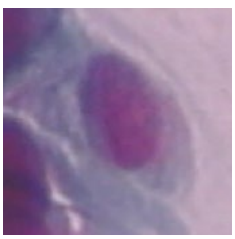
Table 6. Progressive refinement captured at various timesteps on samples from SIPaKMeD.

T= 750 Blob emergence	T=500 Boundary form	T=250 Sharp refinement	T=0 Final output
			
SSIM=0.17	SSIM=0.49	SSIM= 0.90	SSIM= 1.00
			
SSIM = 0.11	SSIM = 0.37	SSIM = 0.85	SSIM = 1.00
			
SSIM = 0.36	SSIM = 0.67	SSIM = 0.92	SSIM = 1.00
			
SSIM = 0.44	SSIM = 0.74	SSIM = 0.94	SSIM = 1.00
			
SSIM = 0.43	SSIM = 0.74	SSIM = 0.95	SSIM = 1.00
			
SSIM = 0.43	SSIM = 0.73	SSIM = 0.95	SSIM = 1.00

At the final step $t = 0$, the segmentation masks closely match the ground truth, with well-defined boundaries, correct topology, and preserved fine-grained structural details. Comparing the progressive refinement across the three visualized cases representing easy (single nucleus, high contrast), moderate (two nuclei with slight overlap), and difficult (irregular morphology with weak staining) scenarios reveals that the diffusion process exhibits robust convergence behavior regardless of case difficulty. Even for the challenging case in Row 3, where initial structure emergence at $t = 750$ is less distinct than in simpler cases, the model successfully refines the segmentation by $t = 0$.

Table 7 shows the progressive refinement of images from the Herlev dataset across T values from 750 to 0, along with the corresponding SSIM scores.

Table 7. Progressive refinement captured at various timesteps on samples from SIPaKMeD.

T= 750 Blob emergence	T=500 Boundary form	T=250 Sharp refinement	T=0 Final output
			
SSIM = 0.50	SSIM = 0.81	SSIM = 0.97	SSIM = 1.00
			
SSIM = 0.28	SSIM = 0.64	SSIM = 0.93	SSIM = 1.00
			
SSIM = 0.47	SSIM = 0.80	SSIM = 0.97	SSIM = 1.00
			
SSIM = 0.51	SSIM = 0.81	SSIM = 0.97	SSIM = 1.00
			
SSIM = 0.60	SSIM = 0.85	SSIM = 0.98	SSIM = 1.00

To quantitatively assess the quality of intermediate predictions, we computed the Structural Similarity Index Measure (SSIM) between outputs at each timestep and the final prediction (t = 0). The overall SSIM values increase monotonically from 0.23 at t = 750 to 0.61 at t = 500, 0.84 at t = 250, and 1.0 at t = 0, confirming progressive improvement in structural coherence. The steepest increase in SSIM occurs between t = 500 and t = 250, corresponding to the phase when boundary refinement is most active. This observation validates our design choice to inject WCRB's boundary features at multiple decoder levels, ensuring that edge information is available precisely when it has maximal impact on boundary sharpening.

4.6. Limitations and Failure Case Analysis

While the proposed method achieves state-of-the-art performance, it is not without limitations. Analysis of the test set reveals that approximately 2.3% of predictions exhibit noticeable errors, primarily occurring in two scenarios. First, in extremely densely packed cell clusters with more than five nuclei in proximity and extensive overlapping regions, the model occasionally merges two adjacent nuclei into a single component. However, such cases represent less than 1% of the test set and are challenging even for expert annotators. Second, in rare cases with severe staining artefacts (e.g., precipitate deposits or uneven dye distribution creating spurious edges), the model may introduce minor boundary irregularities. However, ASNS's noise suppression significantly reduces this issue compared to baselines.

These failure modes suggest directions for future improvement, including the incorporation of instance segmentation techniques to better handle dense clustering and integration of staining normalization as a preprocessing step. Nevertheless, the 97.7% success rate and the clinical acceptability of most errors (boundary deviations within 5 pixels) indicate that the proposed AI tool offers reliable performance for practical cervical cancer screening applications. Future work will explore transformer-based semantic branches and hybrid CNN-Transformer diffusion architectures.

5. Conclusion

This work presents a novel triplet-branch diffusion-based AI tool for cervical nucleus segmentation that addresses boundary ambiguity, noise-induced feature degradation, and topological inconsistency through three synergistic innovations: the Wavelet-Enhanced Contour Refinement Branch, the Adaptive Spectral Noise Suppression y module, and the Topology-Aware Hybrid Loss. Comprehensive evaluation on the Herlev and SIPaKMeD datasets demonstrates superior performance compared to state-of-the-art methods. The model achieves 94.45% Dice, 90.87% IoU, and 7.31 pixels HD95 on Herlev, representing improvements of 2.32%, 2.12%, and 19.2%, respectively, over the best baseline. Similar gains are observed on SIPaKMeD (94.81% Dice, 91.42% IoU, 8.93 pixels HD95). Systematic ablation studies confirm that each component contributes measurably to overall performance, with WCRB providing the most substantial improvements to boundary metrics, ASNS enhancing segmentation accuracy through noise suppression, and TAHL ensuring topological correctness. The clinical significance extends beyond numerical metrics. The 19.2% reduction in boundary localization error directly improves nuclear morphometry measurements, which are critical for grading cervical lesions. The explicit topological constraints prevent anatomically implausible artifacts such as fragmented nuclei, which could trigger false positives in automated diagnostic systems. With approximately 28.7 million parameters and 1.8 seconds of inference time per image, the AI tool is practically deployable in clinical settings. While achieving 97.7% success rate, limitations remain in extremely dense cell clusters (>5 nuclei) and cases with severe staining artefacts. Future work should explore instance segmentation techniques to better handle overlapping structures, domain adaptation for cross-laboratory generalization, and extension to multi-class segmentation that encompasses both nucleus and cytoplasm regions. Prospective clinical validation studies are necessary to assess the tool's impact on diagnostic accuracy in real-world screening workflows. The proposed framework demonstrates that explicit incorporation of domain-specific inductive biases, learnable boundary features, adaptive noise suppression, and topological constraints substantially enhances diffusion-based medical image segmentation, providing a solid foundation for automated cervical cancer screening applications.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] L. Mukku and J. Thomas, "Deep learning-based cervical lesion segmentation in colposcopic images," *Appl. Eng. Technol.*, vol. 3, no. 1, pp. 16–25, Apr. 2024, doi: 10.31763/aet.v3i1.1345.

[3] L. H. Ellenson and T.-C. Wu, "Focus on endometrial and cervical cancer," *Cancer Cell*, vol. 5, no. 6, pp. 533–538, Jun. 2004, doi: 10.1016/j.ccr.2004.05.029.

[4] WHO, "WHO guideline for screening and treatment of cervical pre-cancer lesions for cervical cancer prevention : use of dual-stain cytology to triage women after a positive test for human papillomavirus (HPV)," p. 51, 2024. [online]. Available at: https://www.who.int/publications/i/item/9789240091658.

[5] L. Mukku and J. Thomas, "TelsNet: temporal lesion network embedding in a transformer model to detect cervical cancer through colposcope images," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 3, p. 502, Nov. 2023, doi: 10.26555/ijain.v9i3.1431.

[6] M. Lalasa and J. Thomas, "A Review of Deep Learning Methods in Cervical Cancer Detection," in *Lecture Notes in Networks and Systems*, Springer, Cham, 2023, pp. 624–633, 2023, doi: 10.1007/978-3-031-27524-1_60.

[7] N. B and I. V, "Enhanced machine learning based feature subset through FFS enabled classification for cervical cancer diagnosis," *Int. J. Knowledge-based Intell. Eng. Syst.*, vol. 26, no. 1, pp. 79–89, Jun. 2022, doi: 10.3233/KES-220009.

[8] H. Tang, C. Song, and M. Qian, "Automatic segmentation algorithm for breast cell image based on multi-scale CNN and CSS corner detection," *Int. J. Knowledge-based Intell. Eng. Syst.*, vol. 24, no. 3, pp. 195–203, Sep. 2020, doi: 10.3233/KES-200041.

[9] A. Sahoo and S. Chandra, "Medical image segmentation schemes for the analysis of gynaecological malignancies," *Int. J. Knowledge-based Intell. Eng. Syst.*, vol. 17, no. 4, pp. 291–304, Nov. 2013, doi: 10.3233/KES-130279.

[10] K. Gong, K. Johnson, G. El Fakhri, Q. Li, and T. Pan, "PET image denoising based on denoising diffusion probabilistic model," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 51, no. 2, pp. 358–368, 2024, doi: 10.1007/s00259-023-06417-8.

[11] K. Chen *et al.*, "Quantifying uncertainty: Air quality forecasting based on dynamic spatial-temporal denoising diffusion probabilistic model," *Environ. Res.*, vol. 249, p. 118438, 2024, doi: 10.1016/j.envres.2024.118438.

[12] J. Wu *et al.*, "MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model," *Proceedings of Machine Learning Research*, vol. 227. PMLR, pp. 1623–1639, Jan. 2024. available at: https://proceedings.mlr.press/v227/wu24a.html.

[13] T. Chen, C. Wang, and H. Shan, "BerDiff: Conditional Bernoulli Diffusion Model for Medical Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Cham, Apr. 2023, pp. 491–501. doi: 10.1007/978-3-031-43901-8_47.

[14] C. H. Wong, "EnsDiff: Ensemble Precipitation Nowcasting with Diffusion," *Ensemble Precip. Nowcasting with Diffus.*, no. January, p. 14, 2025. Available at: Google Scholar.

[15] M. Xia *et al.*, "Anatomically and Metabolically Informed Diffusion for Unified Denoising and Segmentation in Low-Count PET Imaging," *Med. Image Anal.*, vol. 107, p. 103831, Oct. 2025, https://doi.org/10.1016/j.media.2025.103831.

[16] M. Xia *et al.*, "Multimodal Spatiotemporal Feature-Based Human Motion Pattern Recognition With CNN-Transformer-Attention Framework," *IEEE Internet Things J.*, vol. 12, no. 20, pp. 43883–43895, 2025, doi: 10.1109/JIOT.2025.3599403.

[17] A. Halder and D. Dey, "MorphAttnNet: An Attention-based morphology framework for lung cancer subtype classification," *Biomed. Signal Process. Control*, vol. 86, p. 105149, 2023, doi: https://doi.org/10.1016/j.bspc.2023.105149.

[18] X. Fan, Y. Lu, B. Hu, Y. Shi, and B. Sun, "LW-MorphCNN: a lightweight morphological attention-based subtype classification network for lung cancer," *Meas. Sci. Technol.*, vol. 36, no. 1, p. 15703, 2025, doi: 10.1088/1361-6501/ad8a7c.

[19] B. Patnaik, D. S. K. Nayak, and S. Sahoo, "Attention enhanced hybrid deep learning model with 1D-CNN and BiLSTM for automated sleep apnea detection," *Discov. Appl. Sci.*, vol. 7, no. 12, p. 1376, 2025, doi: 10.1007/s42452-025-07639-1.

[20] H. A. Phoulady, M. Zhou, D. B. Goldgof, L. O. Hall, and P. R. Mouton, "Automatic quantification and classification of cervical cancer via Adaptive Nucleus Shape Modeling," in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, Sep. 2016, pp. 2658–2662. doi: 10.1109/ICIP.2016.7532841.

[21] A. Gençtav, S. Aksoy, and S. Önder, "Unsupervised segmentation and classification of cervical cell images," *Pattern Recognit.*, vol. 45, no. 12, pp. 4151–4168, Dec. 2012, doi: 10.1016/j.patcog.2012.05.006.

[22] M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering.," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 2, pp. 233–41, Mar. 2011, doi: 10.1109/TITB.2010.2087030.

[23] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, "Stacked Predictive Sparse Decomposition for Classification of Histology Sections," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 3–18, May 2015, doi: 10.1007/s11263-014-0790-9.

[24] T. Chankong, N. Theera-Umpon, and S. Auephanwiriyakul, "Automatic cervical cell segmentation and classification in Pap smears," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 539–556, Feb. 2014, doi: 10.1016/j.cmpb.2013.12.012.

[25] Z. Xing *et al.*, "Diff-UNet: A diffusion embedded network for robust 3D medical image segmentation," *Med. Image Anal.*, vol. 105, p. 103654, 2025, doi: 10.1016/j.media.2025.103654.

[26] A. Pratondo, C.-K. Chui, and S.-H. Ong, "Robust Edge-Stop Functions for Edge-Based Active Contour Models in Medical Image Segmentation," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 222–226, Feb. 2016, doi: 10.1109/LSP.2015.2508039.

[27] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, and D. Yang, "EANet: Iterative edge attention network for medical image segmentation," *Pattern Recognit.*, vol. 127, no. July, p. 108636, Jul. 2022, doi: 10.1016/j.patcog.2022.108636.

[28] P. Kumar, "Diffusion Models and Generative Artificial Intelligence: Frameworks, Applications and Challenges: Pranjal Kumar," *Arch. Comput. Methods Eng.*, vol. 32, no. 7, pp. 4049–4092, 2025, https://doi.org/10.1007/s11831-025-10266-z.

[29] M. Zhang, J. Wu, Y. Ren, J. Yang, M. Li, and A. J. Ma, "Diffusionengine: Diffusion model is scalable data engine for object detection," *Pattern Recognit.*, vol. 171, p. 112141, 2026, doi: 10.1016/j.patcog.2025.112141.

[30] M. J. Ignacio, S. Shin, H. Jin, S. J. Yoo, D. Han, and Y.-G. Kim, "Revisiting U-Net: a foundational backbone for modern generative AI," *Artif. Intell. Rev.*, vol. 59, no. 45, pp. 1–52, 2026, doi: 10.1007/s10462-025-11450-0.

[31] S. Xu, B. Yang, R. Wang, D. Yang, J. Li, and J. Wei, "Single Tree Semantic Segmentation from UAV Images Based on Improved U-Net Network," *Drones*, vol. 9, no. 4. p. 237, 2025, doi: 10.3390/drones9040237.

[32] B. H. Qsim, A. M. Khudhur, D. H. Kadir, and D. M. Saleh, "A Wavelet Shrinkage Mixed with a Single-level 2D Discrete Wavelet Transform for Image Denoising," *Kurdistan J. Appl. Res.*, vol. 9, no. 2, pp. 1–12, 2024, doi: 10.24017/science.2024.2.1.

[33] M. Uddin, Z. Fu, and X. Zhang, "Deepfake face detection via multi-level discrete wavelet transform and vision transformer," *Vis. Comput.*, vol. 41, no. 10, pp. 7049–7061, 2025, doi: 10.1007/s00371-024-03791-8.