

Attn-NucleiSeg: Cross-Attention Fusion with Adaptive Scale Selection for Nuclei Segmentation in HE Images

Rameesha Javed ¹, Nimra Bukhari ¹ and Shabir Hussain ^{2,*}

¹ Department of Computer Science, National College of Business Administration and Economics, Alhamrah University, Rahim Yar khan, 64200, Pakistan

² School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China

*Corresponding author: Shabir Hussain (Email: hussain@zzu.edu.cn; shabir.nicaas@gmail.com)

Received: 01/07/2026, Revised: 17/09/2026, Accepted: 25/09/2026

Abstract—Nuclei instance segmentation in histopathology images is an important yet challenging task for cancer diagnosis and treatment planning. Existing methods often rely on single-encoder architectures that may have difficulty capturing both local morphological details and global contextual information, while fixed receptive fields may not adequately represent nuclei with different sizes and appearances. To address these challenges, we propose Attn-NucleiSeg, an integrated framework that combines three key components: a dual-encoder architecture based on ResNet-50+FPN and Swin Transformer for learning complementary local and global features, a Cross-Modal Attention Fusion (CAFM) module for interaction between CNN and Transformer features, and a Dynamic Receptive Field Selection (DRFS) module that adapts the receptive field using dilation rates of 1, 3, and 5. Experiments on the PanNuke dataset show that Attn-NucleiSeg achieves the highest detection performance, with a precision of 0.86, recall of 0.83, and $F_{1,d}$ of 0.84. It also obtains the highest average PQ across all five evaluated nuclear categories, with scores of 0.431, 0.582, 0.425, 0.150, and 0.576 for Inflammatory, Neoplastic, Connective, Dead, and Epithelial nuclei, respectively. On the MoNuSeg validation set, Attn-NucleiSeg achieves the highest scores among the compared methods, including 0.680 bPQ, 0.827 precision, 0.882 recall, 0.841 $F_{1,d}$, and 0.729 Jaccard. These results indicate that Attn-NucleiSeg can provide consistent nucleus detection and segmentation performance across different histopathology datasets.

Index Terms—Cross-Modal Attention, Deep Learning, Dual-Encoder Network, Dynamic Receptive Field, Histopathology Images, Nuclei Instance Segmentation

I. INTRODUCTION

Histopathological image analysis plays an important role in cancer diagnosis and research because it provides detailed information about the cellular and structural characteristics of tissue [1]. Among the different tissue components, nuclei are particularly important because their shape, size, distribution, and appearance provide useful information about pathological changes. Accurate nuclear instance segmentation and

classification can support quantitative tissue analysis and disease assessment [2]. However, manually identifying and classifying nuclei in high-resolution histopathological images is time-consuming and subject to inter-observer variability. Therefore, reliable automated approaches are needed. Nuclear segmentation and classification remain challenging because nuclei vary substantially in size, shape, texture, and appearance [3]. The problem becomes more difficult when nuclei are densely packed or overlap, making their boundaries difficult to distinguish [4,5]. Variations in tissue staining and image acquisition can further alter the appearance of nuclei. In addition, nuclei from different cell types may have similar visual characteristics, whereas nuclei from the same class can exhibit considerable morphological variation. Therefore, an effective model needs to capture fine cellular details while also considering the surrounding tissue context. CNN-based methods have been widely used for histopathological image analysis because convolutional operations effectively learn local spatial features [6]. Architectures such as U-Net [7], ResUNet++, and HoVer-Net [16] have demonstrated their effectiveness in the segmentation and analysis of cellular structures. More recent deep learning approaches have also explored different strategies for learning informative representations from complex biomedical data [8]. Despite these advances, conventional convolutional networks primarily focus on local receptive fields and may have limited ability to model long-range relationships. This limitation can affect the separation of adjacent nuclei and the discrimination of nuclei with similar local appearances. Transformer-based approaches can capture long-range dependencies through self-attention, but their computational requirements can become substantial when processing high-resolution histopathological images. These limitations motivate the development of methods that effectively combine local and global information while maintaining manageable computational costs. To address these challenges, we propose Attn-NucleiSeg, a framework for



nuclear instance segmentation and classification. The primary contributions of the proposed method are as follows:

- **Dual-Encoder Architecture:** We develop a parallel CNN–Transformer encoder that combines ResNet-50 with FPN and Swin Transformer to capture both fine-grained spatial details and broader contextual information for accurate nuclei segmentation.
- **Cross-Modal Attention Fusion (CAFM):** We introduce a bidirectional cross-attention module that allows CNN and Transformer features to exchange information at multiple scales. Rather than simply refining features within a single branch, CAFM enables the two branches to interact and combines their complementary local and global representations through learnable feature fusion.
- **Dynamic Receptive Field Selection (DRFS):** We design a gated module that dynamically selects among dilation rates of 1, 3, and 5 at each spatial location. This allows the network to adjust its receptive field according to local feature characteristics, helping it capture nuclei with different sizes and morphologies without relying on manually selected scales.

II. RELATED WORK

A. CNN-Based Nucleus Segmentation

CNN-based models have been widely used for nucleus segmentation because they can effectively learn detailed spatial features from histopathological images. U-Net introduced the encoder–decoder architecture [7], which has become a foundation for many biomedical image segmentation methods. Several variants have since been developed to improve its segmentation performance. For example, Attention U-Net [9] uses attention gates to focus on relevant regions, while Micro-Net [10] employs multi-resolution feature learning to improve the localization of cellular structures in microscopy images. More recently, WaveSeg-UNet [11] has been proposed for segmenting overlapping nuclei in multi-organ histopathology images. Despite these advances, CNNs mainly capture local spatial information and have limited ability to model long-range dependencies. This limitation can become more apparent in complex and densely packed tissue regions, where separating adjacent nuclei with similar appearances remains challenging.

B. Instant-Aware Nucleus Segmentation

CLA-UNet [12] explored multi-scale feature extraction to better handle differences in nuclear size, shape, and appearance. However, separating individual nuclei remains difficult when cells are closely packed or overlap with each other. Conventional semantic segmentation methods can identify the overall nuclear region but may not clearly separate neighboring nuclei. To overcome this issue, SAM-HoVNet [13] extended the HoVer-Net framework by incorporating the Segment Anything Model (SAM), Squeeze-and-Excitation modules, and multi-scale feature representations for multi-class nucleus segmentation and classification. These approaches show how

combining multi-scale features with instance-level information can improve the separation of adjacent nuclei.

C. Transformer-Based Nucleus Segmentation

Vision Transformers have gained increasing attention in medical image segmentation because self-attention enables models to capture relationships between distant regions of an image. Attention-based approaches have also been explored to incorporate contextual information into medical image analysis. TransUNet [14] combines Transformer-based representations with the U-Net architecture to capture both global and local features, while Swin-Unet [15] uses hierarchical window-based attention to model contextual information more efficiently. In histopathological image analysis, CellViT [20] has further demonstrated the potential of Transformer-based architectures for cell segmentation and classification. Despite their ability to capture global context, Transformer-based models generally require more computational resources than conventional CNNs, while accurately preserving fine nuclear boundaries remains challenging.

III. METHODOLOGY

This section presents the proposed Attn-NucleiSeg framework, detailing its architectural components, novel modules, and training strategy.

A. Dataset

We evaluate Attn-NucleiSeg on two widely used histopathology datasets. PanNuke [18] is used for training and primary evaluation, while MoNuSeg [19] is used to assess cross-dataset generalization. PanNuke contains diverse tissue types and nucleus categories, making it a challenging benchmark for nucleus segmentation and classification. Following [15], we use 14 MoNuSeg test images from different organs for segmentation evaluation, as the dataset does not provide nucleus type annotations.

B. Proposed Framework

The proposed Attn-NucleiSeg in Fig. 1 adopts a U-Net-inspired architecture with a parallel dual-encoder design that combines CNN and Vision Transformer (ViT) backbones to capture complementary local spatial details and global contextual information from H&E images. The network incorporates two key components, Cross-Modal Attention Fusion (CAFM) and Dynamic Receptive Field Selection (DRFS), to further improve feature representation and nuclei segmentation. CAFM performs bidirectional cross-attention between CNN and ViT features at corresponding scales, allowing the two branches to exchange complementary information before their features are combined through learnable weighted fusion. In contrast to conventional attention mechanisms that mainly refine features within a single representation, CAFM is designed to explicitly facilitate information exchange between the CNN and ViT feature streams. The DRFS module complements this cross-modal fusion by addressing the substantial variation in nuclei size and morphology. Rather than relying on fixed receptive fields or predefined multi-scale configurations, DRFS uses multiple dilated convolutions with dilation rates of 1, 3, and 5

together with spatial gating to dynamically select appropriate receptive-field information at each spatial location. This allows the network to capture both small and large nuclear structures without requiring manual selection of a fixed scale. The decoder follows a multi-task design with three prediction branches for nucleus prediction (NP), horizontal-vertical (HV) distance maps, and nucleus type (NT) classification. These outputs are subsequently processed using a HoVer-Net [16] inspired post-processing strategy, including centroid detection, watershed-based instance separation, and majority voting for nucleus-type assignment. This process generates individual nucleus instances together with their corresponding classifications.

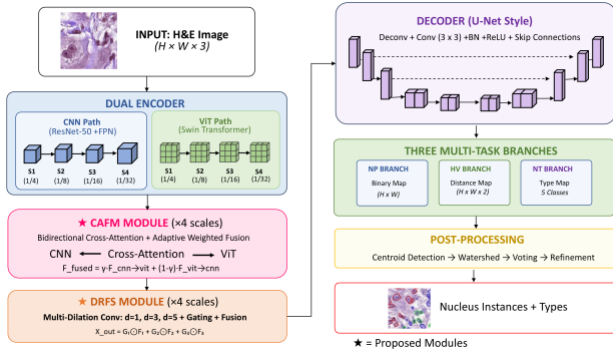


Fig. 1. Overall architecture of Attn-NucleiSeg. The framework employs a dual-encoder with CNN (ResNet-50+FPN) and ViT (Swin Transformer) branches. The two proposed modules, CAFM for cross-attention fusion and DRFS for adaptive scale selection, are integrated at each encoder level. The decoder produces three multi-task outputs: Nucleus Prediction (NP), Horizontal-Vertical distance maps (HV), and Nucleus Type classification (NT). Post-processing yields final instance segmentation with type labels.

C. Dual Encoder

Given an input H&E image $X \in \mathbb{R}^{H \times W \times 3}$, the CNN branch, based on ResNet-50 with FPN, extracts fine-grained local spatial features, while the Swin Transformer branch captures broader contextual information. Both branches generate multi-scale feature representations at four different resolutions, providing complementary local and contextual information for subsequent feature fusion. The extracted features at each scale are represented as

$$F_{cnn}^l = E_{cnn}^l(X), F_{vit}^l = E_{vit}^l(X), l \in \{1, 2, 3, 4\} \quad (1)$$

where E_{cnn}^l and E_{vit}^l denote the CNN and ViT encoders at the l -th scale, respectively in (1).

D. Cross-Model Attention Fusion Module (CAFM)

The CAFM performs bidirectional cross-attention between CNN and ViT features at each scale. The CNN-to-ViT direction enriches local CNN features with global context in (2), while the ViT-to-CNN direction refines global understanding with precise spatial details in (3). The two cross-attended features are defined as

$$F_{c \rightarrow v}^l = \text{MHA}(Q_{cnn}^l, K_{vit}^l, V_{vit}^l) \quad (2)$$

$$F_{v \rightarrow c}^l = \text{MHA}(Q_{vit}^l, K_{cnn}^l, V_{cnn}^l) \quad (3)$$

The final fused feature is obtained via learnable weighted fusion refers to (4):

$$F_{fused}^l = \gamma_l F_{c \rightarrow v}^l + (1 - \gamma_l) F_{v \rightarrow c}^l, \gamma_l \in [0, 1] \quad (4)$$

Here, γ_l is a learnable scale-specific parameter that controls the relative contribution of the two cross-attention directions. MHA denotes multi-head attention with $h=8$, which captures diverse feature interactions between the CNN and ViT representations. The attention operation follows the standard scaled dot-product formulation.

E. Dynamic Receptive Field Selection (DRFS) Module

Cell nuclei vary considerably in size across tissue types. Lymphocytes appear as small, compact structures (5-10 pixels), while carcinoma cells can be large and irregular (30-50+ pixels). Using fixed receptive fields forces a compromise that may fail for both extremes. The DRFS module addresses this by enabling per-pixel adaptive selection of receptive field size based on local nucleus characteristics. It uses three parallel dilated convolutions in (5) with dilation rates $d \in \{1, 3, 5\}$ to capture small-, medium-, and large-scale features:

$$F_d^l = \text{Conv}_{3 \times 3}^d(F_{fused}^l), d \in \{1, 3, 5\} \quad (5)$$

Equation (6) represents a lightweight spatial gating network assigns location-specific weights to these scales based on local features:

$$G^l = \sigma(W_g[F_1^l, F_2^l, F_3^l + b_g]) \quad (6)$$

The resulting weighted fusion enables the network to adapt smoothly to nuclei of different sizes and morphological characteristics in (7):

$$X_{DRFS}^l = \sum_{d \in \{1, 3, 5\}} G_d^l \odot F_d^l \quad (7)$$

Here, σ denotes sigmoid activation, $[\cdot]$ represents channel-wise concatenation, and $\odot$ denotes element-wise multiplication. This design allows the network to adapt to varying nucleus sizes without manual parameter tuning, improving segmentation accuracy across diverse tissue types.

F. Multi-Task Decoder

The decoder progressively upsamples the features from the DRFS module in (7) and integrates multi-scale information through skip connections enhanced by CAFM. It follows a U-Net-style architecture and contains three parallel prediction branches.

Equation (8) represents NP branch generates a binary nucleus map:

$$\hat{Y}_{np} \in [0, 1]^{H \times W} \quad (8)$$

The HV branch predicts horizontal and vertical distance maps for separating adjacent nuclei, while the NT branch classifies nuclei into five tissue-related categories. Together, these branches provide complementary information for accurate nucleus segmentation and classification.

G. Training Strategy

The network is trained using a combined loss that jointly supervises the NP, HV, and NT branches. The total training objective is defined as shown in (9):

$$\mathcal{L}_{total} = \mathcal{L}_{NP} + \mathcal{L}_{HV} + \mathcal{L}_{NT} \quad (9)$$

Equation (10) represents NP and NT branches use combinations of Focal Tversky, Dice, and boundary/edge losses:

$$\begin{aligned} \mathcal{L}_{NP} &= \lambda_1 \mathcal{L}_{FT} + \lambda_2 \mathcal{L}_{DICE} + \lambda_3 \mathcal{L}_{edge}, \\ \mathcal{L}_{NT} &= \lambda_4 \mathcal{L}_{FT} + \lambda_5 \mathcal{L}_{DICE} + \lambda_6 \mathcal{L}_{edge} \end{aligned} \quad (10)$$

The Focal Tversky loss helps address class imbalance, while the Dice and boundary terms improve segmentation and localization. The HV branch is optimized using MSE and multi-scale gradient losses to improve the localization and separation of adjacent nuclei. The loss weights $\lambda_1, \dots, \lambda_6$ are applied according to the training configuration.

For training, the input H&E images are resized to 256×256 pixels and represented as RGB images with three channels. The experiments followed the three predefined folds provided by the PanNuke dataset organizers. The proposed model was trained and evaluated independently on each fold, and the final results were obtained by averaging the performance across all three folds. The model is optimized using AdamW with an initial learning rate of 1×10^{-4} , a momentum parameter of 0.99, and a weight decay of 0.05. A cosine annealing learning-rate schedule is employed with linear warm-up during the first 50 epochs. The network is trained for 100 epochs with a batch size of 16. Data augmentation includes flipping, rotation, scaling, color jittering, Gaussian blur, and elastic deformation. The implementation uses PyTorch 1.11.0 with Python 3.7.0 and CUDA 12.2, and experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory.

H. Post-Processing

Following HoVer-Net [16], post-processing is applied to the predicted horizontal and vertical distance maps to obtain individual nucleus instances. Centroid detection is first performed using the predicted HV maps to identify the centers of nuclei. The detected centroids are then used to guide a watershed algorithm for separating adjacent and overlapping nuclei. For each resulting nucleus instance, the predicted nuclear-type probabilities are aggregated, and majority voting is used to assign the final tissue-related category. Finally, spurious or invalid instances are removed to obtain the final instance-level segmentation and classification results.

IV. RESULTS AND DISCUSSION

In this section, we present and analyze the experimental results obtained using the proposed Attn-NucleiSeg framework. The evaluation focuses on its ability to accurately detect and segment nuclei in histopathology images while maintaining reliable classification performance. We compare the proposed approach with existing state-of-the-art methods using several quantitative evaluation metrics. In addition, we examine the generalization capability of Attn-NucleiSeg across different experimental settings to assess its robustness and consistency.

The results are further analyzed to highlight the effectiveness of the proposed architectural components and their contribution to overall performance.

A. Evaluation Metrics

Following the CellViT evaluation protocol [20], we use Panoptic Quality (PQ) as the primary metric, reporting bPQ for overall nucleus segmentation and mPQ for joint segmentation and classification. Detection performance is evaluated using P_d , R_d and $F_{1,d}$, where d represents binary nucleus detection across all classes. We additionally report P_c , R_c , and $F_{1,c}$ for each of the five nucleus types to assess class-specific performance.

B. Nucleus Detection Performance on PanNuke

The nucleus detection results of Attn-NucleiSeg on the PanNuke dataset are presented in Table I, using precision (P_d), recall (R_d), and F1-score ($F_{1,d}$) as the evaluation measures. Attn-NucleiSeg obtains a precision of 0.86 and a recall of 0.83, indicating that the model is able to detect nuclei while maintaining a good balance between correct and missed detections. These values are also reflected in its $F_{1,d}$ score of 0.84, which is higher than the scores reported for CellViT (SAM) (0.83), STARDIST (0.82), CellViT (ViT256) (0.82), HoVer-Net (0.80), and Mask-RCNN (0.72). In particular, the improvement in recall over STARDIST and HoVer-Net [16] suggests that Attn-NucleiSeg is better able to identify nuclei across the varied cellular patterns present in PanNuke, while the higher precision indicates fewer incorrect detections. These results demonstrate consistent nucleus localization performance across different tissue types and cellular appearances. This performance can be attributed, in part, to the complementary feature representations learned through the dual-encoder architecture, which captures both local and broader contextual information. In addition, the Cross-Modal Attention Fusion Module (CAFM) and Dynamic Receptive Field Selection (DRFS) help the network focus on relevant spatial features at different scales during feature extraction. Together, these components provide more informative representations for distinguishing individual nuclei, particularly in regions with variations in nuclear size, shape, and density.

TABLE I
THE DETECTION PERFORMANCE OF ATTN-NUCLEISEG ON THE PANNUKE DATASET.

Model	Detection		
	P_d	R_d	$F_{1,d}$
CellViT(SAM) [20][17]	0.83	0.82	0.83
Mask-RCNN [22]	0.76	0.68	0.72
STARDIST [23]	0.85	0.8	0.82
HoVer-Net [16]	0.82	0.79	0.8
CellViT(ViT256) [20][21]	0.83	0.82	0.82
Attn-NucleiSeg (Ours)	0.86	0.83	0.84

C. Nucleus Classification Performance on PanNuke

The classification results of Attn-NucleiSeg on the PanNuke dataset are reported in Table II for five nucleus categories: Dead, Neoplastic, Inflammatory, Connective, and Epithelial. The proposed model shows variation in classification performance across the five categories, reflecting differences in their visual appearance and morphological characteristics. For Dead nuclei, the model achieves an F₁-score of 0.38, with a precision of 0.45 and recall of 0.39. The relatively lower performance of this category may be associated with the heterogeneous and less well-defined appearance of dead nuclei, which can exhibit weak nuclear boundaries, fragmented structures, and variable staining characteristics. These features can make them difficult to distinguish from surrounding tissue and other nucleus types. The Neoplastic category achieves a higher F₁-score of 0.71, with precision and recall of 0.72 and 0.71, respectively. For Inflammatory nuclei, the model obtains an F₁-score of 0.61, with precision of 0.61 and recall of 0.59. The Connective category achieves an F₁-score of 0.55, with precision and recall of 0.56 and 0.53, respectively. The comparatively lower performance for Connective nuclei may be related to their morphological variability and visual similarity to other elongated, irregular, or densely distributed nuclei in the surrounding tissue. In addition, the appearance of connective-tissue-associated nuclei can vary considerably across different tissue regions, making consistent class-specific feature learning more challenging. In contrast, the Epithelial category achieves the highest classification performance, with an F₁-score of 0.73, precision of 0.75, and recall of 0.74, indicating more consistent discriminative features for this category. Overall, the results indicate that categories with more distinctive and consistent morphological patterns are easier to classify, whereas Dead and Connective nuclei remain more challenging due to their heterogeneous appearance and potential visual overlap with other categories. The dual-encoder architecture, together with the Cross-Modal Attention Fusion Module (CAFM) and Dynamic Receptive Field Selection (DRFS), provides complementary local, contextual, and multi-scale features that can support classification across these diverse nucleus categories; however, the remaining differences in performance indicate that distinguishing visually ambiguous classes remains challenging.

D. Nucleus Segmentation Performance on PanNuke

The segmentation performance of Attn-NucleiSeg is assessed on the PanNuke dataset using Panoptic Quality (PQ). The values reported in Table III correspond to the average PQ obtained across the three PanNuke splits for each nuclear category. For comparison, the proposed method is evaluated against HoVer-Net [16], STARDIST [25], CellViT (ViT256) [20][23], and CellViT (SAM) [20][17]. The ViT256 and SAM versions of CellViT use ViT256 and SAM [17], respectively, as their pre-trained models. As shown in Table III, Attn-NucleiSeg gives the highest average PQ for all five nuclear categories. The model achieves 0.431 for Inflammatory nuclei, 0.582 for Neoplastic nuclei, 0.425 for Connective nuclei, 0.150 for Dead nuclei, and 0.576 for Epithelial nuclei. Among the compared methods, CellViT (SAM) generally provides the closest results,

TABLE II
CLASSIFICATION PERFORMANCE OF ATTN-NUCLEISEG ON THE PANNUKE DATASET.

Methods	Classification														
	Dead			Neoplastic			Inflammatory			Connective			Epithelial		
	F _{1,Dead}	R _{Dead}	P _{Dead}	F _{1,Neo}	R _{Neo}	P _{Neo}	F _{1,Inf}	R _{Inf}	P _{Inf}	F _{1,Con}	R _{Con}	P _{Con}	F _{1,Epi}	R _{Epi}	P _{Epi}
CellViT(SAM) [20][17]	0.36	0.29	0.46	0.70	0.70	0.70	0.59	0.58	0.59	0.53	0.52	0.54	0.71	0.72	0.71
Mask-RCNN [22]	0.22	0.30	0.17	0.59	0.63	0.55	0.50	0.54	0.46	0.42	0.43	0.42	0.52	0.52	0.52
STARDIST [23]	0.10	0.09	0.39	0.69	0.69	0.69	0.57	0.53	0.62	0.51	0.49	0.54	0.70	0.68	0.73
HoVer-Net [16]	0.31	0.35	0.28	0.62	0.67	0.58	0.54	0.51	0.56	0.49	0.47	0.52	0.56	0.60	0.54
CellViT(ViT256) [20][21]	0.37	0.35	0.39	0.69	0.70	0.69	0.58	0.58	0.59	0.52	0.51	0.53	0.70	0.71	0.68
Attn-NucleSeg (Ours)	0.38	0.39	0.45	0.71	0.71	0.72	0.61	0.59	0.61	0.55	0.53	0.56	0.73	0.74	0.75

with values of 0.425, 0.574, 0.415, and 0.565 for Inflammatory, Neoplastic, Connective, and Epithelial nuclei, respectively. The proposed method also performs slightly better for the Dead category, reaching 0.150 compared with 0.144 for CellViT (ViT256) and 0.140 for CellViT (SAM). The improvements are consistent across the categories rather than being limited to a particular nucleus type. Overall, these results demonstrate that Attn-NucleiSeg maintains reliable segmentation quality across the different nuclear categories in PanNuke. This performance may be supported by the dual-encoder architecture, where CAFM and DRFS help the network retain complementary spatial and contextual information during the segmentation process.

TABLE III
COMPARISON OF AVERAGE PQ ACROSS NUCLEAR CATEGORIES ON THE PANNUKE DATASET

Model	Inflam- matory	Neo- plastic	Connec- tive	Dead	Epi- thelial
CellViT (SAM) [20][17]	0.425	0.574	0.415	0.14	0.565
Mask R- CNN [22]	0.29	0.472	0.3	0.069	0.403
STARDIST [23]	0.398	0.564	0.388	0.024	0.543
HoVer-Net [16]	0.417	0.551	0.388	0.139	0.491
CellViT (ViT256) [20][21]	0.405	0.567	0.405	0.144	0.559
Attn- NucleiSeg (Ours)	0.431	0.582	0.425	0.15	0.576

E. Qualitative Visualization Results on PanNuke

The qualitative segmentation results of Attn-NucleiSeg are shown for five different tissue types from the PanNuke dataset. The Fig. 2 presents the input image patches together with their ground-truth annotations and the corresponding predictions from the proposed model. Across the different tissue types, the predicted masks generally follow the annotated nuclear regions and preserve the shapes and boundaries of individual nuclei. The model also maintains reasonable separation between closely located nuclei, including areas with dense and varied nuclear structures. In addition, the predictions capture nuclei with different sizes and morphological appearances across the selected tissue samples. Some variation can be observed in complex regions where nuclei are closely packed or have irregular boundaries, but the overall segmentation remains consistent with the ground-truth annotations. The visual comparison shows that the model preserves the structure and boundaries of individual nuclei while avoiding excessive merging and fragmentation. The predicted masks closely follow the nuclei in the original histopathological images, even in areas where the nuclei are densely packed or partially overlapping. These observations provide a qualitative view of the segmentation capability of Attn-NucleiSeg and indicate its ability to produce clear and well-separated nuclear regions.

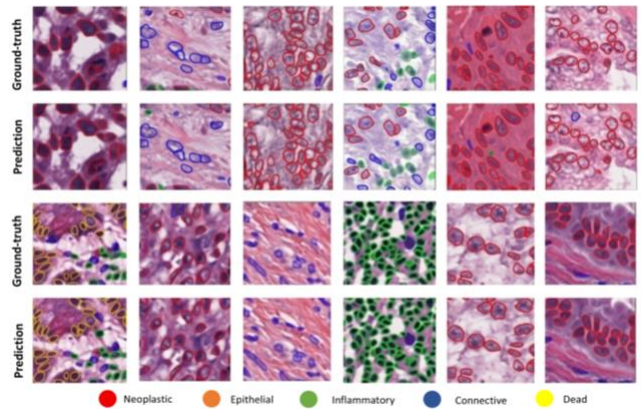


Fig. 2. Qualitative visualization of Attn-NucleiSeg predictions across five tissue types in the PanNuke dataset, with ground-truth annotations and predicted segmentation masks overlaid on the corresponding image patches.

F. Nucleus Detection and Segmentation Performance on MoNuSeg

The results obtained on the MoNuSeg validation set are given in Table IV. Attn-NucleiSeg achieves the highest scores for all the evaluated metrics, with 0.680 bPQ, 0.827 precision, 0.882 recall, 0.841 $F_{1,d}$, and 0.729 Jaccard. The recall shows a clear improvement over both CellViT (ViT256) (0.786) and CellViT (SAM) (0.866), indicating that the proposed method is able to detect a larger number of nuclei correctly. At the same time, its precision of 0.827 remains comparable to the other methods, showing that this improvement in recall does not come at the cost of a large number of false detections. The proposed method also obtains better bPQ and Jaccard scores than both comparison methods, with improvements from 0.664 to 0.680 and from 0.712 to 0.729 over CellViT (ViT256), respectively. Overall, these results show that Attn-NucleiSeg provides more consistent detection and segmentation performance on the MoNuSeg validation set.

TABLE IV
 $F_{1,d}$ -BASED NUCLEUS DETECTION RESULTS ON THE MONUSEG VALIDATION SET.

Method	bPQ	P_d	R_d	$F_{1,d}$	Jaccard
CellViT (ViT256) [20][21]	0.664	0.824	0.786	0.804	0.712
CellViT (SAM) [20] [17]	0.668	0.809	0.866	0.836	0.710
Attn-NucleiSeg (Ours)	0.680	0.827	0.882	0.841	0.729

G. Qualitative Visualization Results on MoNuSeg

The qualitative segmentation results on the MoNuSeg dataset are presented in Fig. 3. The figure compares the predicted nuclear regions with their corresponding ground-truth annotations for representative tissue images. The selected samples contain nuclei with different sizes, shapes, and spatial arrangements, allowing the segmentation performance to be observed under varied tissue conditions. Overall, Attn-NucleiSeg produces predictions that closely follow the boundaries of the annotated nuclei and maintains good separation between neighboring nuclei, even in densely packed

regions. The predicted masks also preserve the overall shape and structure of individual nuclei with relatively good agreement with the ground truth. These visual observations are consistent with the quantitative results obtained on the MoNuSeg validation set and further demonstrate the effectiveness of Attn-NucleiSeg for segmenting nuclei with diverse morphological characteristics.

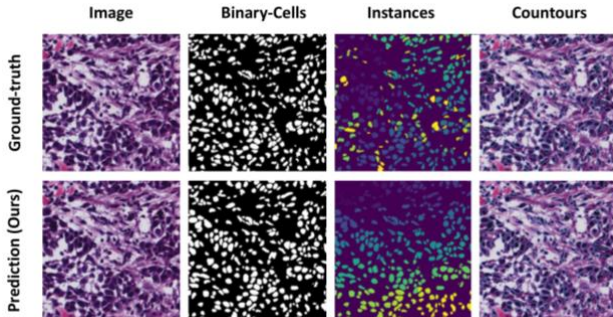


Fig. 3. Qualitative visualization of Attn-NucleiSeg predictions on representative MoNuSeg images with corresponding ground-truth annotations.

H. Ablation Study

We perform ablation studies on PanNuke to examine the contribution of the proposed components in Table V. The experiments progressively add the dual encoder, Cross-Modal Attention Fusion (CAFM), and Dynamic Receptive Field Selection (DRFS), with performance evaluated using bPQ and mPQ. The baseline achieves 0.659 bPQ and 0.478 mPQ. Adding the dual encoder, consisting of the ResNet-50 with FPN CNN branch and Swin Transformer branch, improves the results to 0.667 bPQ and 0.485 mPQ. The addition of CAFM further increases the scores to 0.674 and 0.496, respectively, by

TABLE V
ABLATION STUDY OF ATTN-NUCLEISEG ON THE PANNUKE DATASET, SHOWING THE CONTRIBUTIONS OF THE DUAL ENCODER, CAFM, AND DRFS USING BPQ AND MPQ.

Configuration	Dual Encoder	CAFM	DRFS	bPQ	mPQ
Baseline	X	X	X	0.659	0.478
+ Dual Encoder	✓	X	X	0.667	0.485
+ CAFM	✓	✓	X	0.674	0.496
Attn-NucleiSeg (Proposed)	✓	✓	✓	0.680	0.504

enabling bidirectional interaction between CNN and ViT features. Finally, incorporating DRFS, which provides adaptive receptive fields for nuclei of different sizes, results in the complete Attn-NucleiSeg model with 0.680 bPQ and 0.504 mPQ. The consistent improvement across the configurations

indicates the contribution of each proposed component to the overall performance.

V. CONCLUSION

In this work, we proposed Attn-NucleiSeg, an integrated framework for nucleus detection, instance segmentation, and classification in histopathology images. The dual-encoder architecture combines local morphological and global contextual features, while CAFM and DRFS help the network learn features from nuclei with different shapes, sizes, and appearances. Experiments on PanNuke show that Attn-NucleiSeg provides consistent performance across the evaluated detection, classification, and segmentation tasks. The results on MoNuSeg also show competitive performance across the evaluated metrics, indicating that the method can be applied to a different histopathology dataset. The qualitative results further show reasonable agreement between the predicted masks and ground-truth annotations. Overall, the findings suggest that Attn-NucleiSeg provides a practical approach for automated nuclear analysis in histopathology images. Future work will consider evaluation on more diverse datasets and further improvement of segmentation and classification for challenging nuclear categories.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

AUTHOR CONTRIBUTIONS

Conceptualization, R.J. and N.B.; methodology, R.J. and N.B.; software, N.B.; validation, N.B.; formal analysis, R.J. and N.B.; investigation, S.H.; resources, S.H.; data curation, N.B.; writing—original draft preparation, R.J.; writing—review and editing, S.H.; supervision, S.H.; project administration, S.H.

FUNDING STATEMENT

This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

CONSENT FOR PUBLICATION

All authors have read and approved the final version of the manuscript and consent to its publication.

DECLARATION OF GENERATIVE AI IN SCIENTIFIC WRITING

For preparing this manuscript, the authors have only made use of generative artificial intelligence systems to enhance language proficiency, grammar, and readability.

REFERENCES

- [1] Bukhari, N., Munir, H., Haider, R., & Hussain, S. MASKYLO: Hybrid Deep Learning Framework for Detection and Segmentation of HE Stained Histology Image. *Pakistan Journal of Scientific Research*, 4(2 (Suppl.)), 92-101, 2025.
- [2] Gulsen, A. Benchmarking Deep Segmentation Architectures for Histopathological Nuclei Segmentation: A Controlled Study of Transfer Learning, Robustness, and Efficiency Across CNN, Transformer, and Hybrid Models. *Journal of Imaging Informatics in Medicine*, 1-21, 2026.
- [3] Rasheed, A., Shirazi, S. H., Khan, P., Aseere, A. M., & Shahzad, M. Techniques and challenges for nuclei segmentation in cervical smear images: a review. *Artificial Intelligence Review*, 58(10), 295, 2025.
- [4] Debsarkar, S. S., Aronow, B., & Prasath, V. S. Advancements in automated nuclei segmentation for histopathology using you only look once-driven approaches: A systematic review. *Computers in biology and medicine*, 190, 110072, 2025.
- [5] Hussain, S., Ghaffar, M., & Babar, R. Overcoming Temporal and Sequential Data Challenges in Electroencephalography for Harmful Brain Activity Classification. *Biomed. Eng. Commun.*, 5, 8, 2026.
- [6] Bukhari, N., Hussain, S., Shahzada, H. M. A., Haider, R., Munir, H., Mudassar, I. A., ... & Khan, A. Multi-scale dual-stream cnn-mamba hybrid framework for accurate nuclei segmentation in h&e images. *Pakistan Journal of Scientific Research*, 5(1), 2025.
- [7] Ronneberger, O., Fischer, P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015, (pp. 234-241). Cham: Springer international publishing.
- [8] Hussain, S., Ayoub, M., Ye, F., & Liu, X. MLN2SVG: domain-aware spatially variable gene detection using contrastive variational autoencoder and multi-level neighbor search. *Briefings in Bioinformatics*, 27(3), bbag210, 2026.
- [9] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [10] S. E. A. Raza, L. Cheung, M. Shaban, S. Graham, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Micro-net: A unified model for segmentation of various objects in microscopy images," *Medical Image Analysis*, vol. 52, pp. 160–173, 2019.
- [11] Khan, H. U., Raza, B., Khan, M. A. I., & Faheem, M. WaveSeg-UNet model for overlapped nuclei segmentation from multi-organ histopathology images. *CAAI Transactions on Intelligence Technology*, 10(1), 253-267, 2025.
- [12] Guo, W., Liu, Z., & Ou, Y. CLA-UNet: Convolution and Focused Linear Attention Fusion for Tumor Cell Nucleus Segmentation. *International Journal of Imaging Systems and Technology*, 35(2), e70041, 2025.
- [13] Liu H, Xu H, Pölönen I, Tang C, Zhang M, Yao S, Lu C, Cong F. EMBC Special Issue: SAM-HovNet: SAM-Enhanced Multi-Class Nucleus Segmentation and Classification in Hepatocellular Carcinoma Slides. *IEEE Transactions on Biomedical Engineering*. 2026 Aug 4. doi: 10.1109/TBME.2026.3720392.
- [14] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. 2021 Feb 8. *arXiv preprint arXiv:2102.04306*.
- [15] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision, 2022*, (pp. 205-218). Cham: Springer Nature Switzerland.
- [16] Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58, 101563, 2019.
- [17] Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. Segment anything in medical images. *Nature communications*, 15(1), 654, 2024.
- [18] J. Gamper, N. A. Koohbanani, K. Benes, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," 2020.
- [19] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, and e. a. Heng, Pheng-Ann, "A multi-organ nucleus segmentation challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1380–1391, 2020.
- [20] Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., ... & Kleesiek, J. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical image analysis*, 94, 103143, 2024.
- [21] Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16123-16134). IEEE.
- [22] Abdulla, W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow, 2017.
- [23] S. Chen, C. Ding, M. Liu, J. Cheng, and D. Tao, "Cpp-net: Context aware polygon proposal network for nucleus segmentation," *IEEE Trans actions on Image Processing*, vol. 32, pp. 980–994, 2023.