Alexandria Engineering Journal 110 (2025) 557–566

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Original article

MoE-NuSeg: Enhancing nuclei segmentation in histology images with a two-stage Mixture of Experts network

Xuening Wu [a], Yiqing Shen [b], Qing Zhao [a], Yanlan Kang [a], Wenqiang Zhang [c,*]

[a] Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University, Shanghai 200433, China
[b] Department of Computer Science, Johns Hopkins University, Baltimore 21218, USA
[c] Engineering Research Center of AI & Robotics, Ministry of Education, Academy for Engineering & Technology, School of Computer Science, Fudan University, Shanghai 200433, China

ARTICLE INFO

Keywords:
Nuclei segmentation
Histology image
Swin transformer
Mixture of Experts (moEs)

ABSTRACT

Accurate nuclei segmentation is essential for extracting quantitative information from histology images to support disease diagnosis and treatment decisions. However, precise segmentation is challenging due to the presence of clustered nuclei, varied morphologies, and the need to capture global spatial correlations. While state-of-the-art Transformer-based models employ tri-decoder architectures to decouple the segmentation task into nuclei, edges, and cluster edges segmentation, their complexity and long inference times hinder clinical integration. To address this, we introduce MoE-NuSeg, a novel Mixture of Experts (MoE) network that consolidates the tri-decoder into a single decoder. MoE-NuSeg employs three specialized experts for nuclei segmentation, edge delineation, and cluster edge detection, thereby mirroring the functionality of tri-decoders while surpassing their performance and reducing parameters by sharing attention heads. We propose a two-stage training strategy: the first stage independently trains the three experts, and the second stage fine-tunes their interactions to dynamically allocate the contributions of each expert using a learnable attention-based gating network. Evaluations across three datasets demonstrate that MoE-NuSeg outperforms the state-of-the-art methods, achieving an average increase of 0.99% in Dice coefficient, 1.14% in IoU and 0.92% in F1 Score, while reducing parameters by 30.1% and FLOPs by 40.2%. The code is available at https://github.com/deep-geo/MoE-NuSeg.

1. Introduction

Nuclei segmentation is a fundamental step in assessing nuclei features. These features can help identify abnormal cells or tissues, aiding in cancer diagnosis, grading, and the development of treatment strategies [1]. However, manual annotation of nuclei contours remains a laborious and time-consuming task, prone to inter-observer variability and human error [2]. The advent of whole-slide imaging (WSI) and the evolution of deep learning approaches have facilitated the automation of nuclei segmentation by providing high-resolution digital images and enabling the development of advanced segmentation methods [3,4].

In the early stages of deep learning, Convolutional Neural Networks (CNNs) gained prominence, particularly with their multi-layered encoder–decoder U-shaped architectures, pioneered by U-Net [5], which later inspired numerous variants [6–8], as depicted in Fig. 1(a). CNNs are favored over handcrafted feature-based approaches due to their ability to learn complex features directly from data, without the need for manual feature engineering. They also offer robustness to variations in histology image quality and noise, and have demonstrated strong performance in end-to-end training [9,10].

One major challenge in nuclei segmentation is the accurate separation of clustered nuclei, which most CNN-based methods struggled to address [11]. This difficulty arises partly from the visual similarity between clustered nuclei and the lack of clear boundaries separating them [1]. To tackle these challenges, previous works have explored adapting domain knowledge for precise nuclear boundary delineation and the separation of clustering or overlapping nuclei by designing advanced architectures. For instance, the deep contour-aware neural network (DCAN) [12] decouples the nuclei segmentation task into two subtasks: nuclei detection and edge detection, enabling the model to learn more specialized and discriminative features for each subtask. As illustrated in Fig. 1(b) and (c), this approach has evolved from single-decoder architectures to more sophisticated configurations, including bi-decoder [12,13], intermediate 2.5-decoder [14], and ultimately, tri-decoder architectures [15].

* Corresponding author.
E-mail addresses: yshen92@jhu.edu (Y. Shen), wqzhang@fudan.edu.cn (W. Zhang).

https://doi.org/10.1016/j.aej.2024.10.011
Received 12 July 2024; Received in revised form 19 September 2024; Accepted 3 October 2024
Available online 16 October 2024

However, as the number of decoders increases, the number of parameters and computational demands scale accordingly [11,14,15]. Moreover, multi-decoder models tend to learn redundant feature representations across decoders, which reduces overall efficiency.

In addition to their inductive bias toward locality [16], CNN-based models often struggle to handle varying scales and complex shapes of nuclei due to their reliance on fixed-size receptive fields. This inherent limitation restricts their ability to adapt to diverse nuclei sizes and shapes, which is essential for accurately capturing global spatial relationships and internuclear correlations [17]. As a result, this rigidity leads to suboptimal performance in detecting and segmenting irregularly shaped or closely packed nuclei.

To address these challenges, we propose MoE-NuSeg, a Mixture of Experts (MoE) network [18] based on the Swin Transformer [19], specifically designed for nuclei segmentation. MoE-NuSeg simplifies the tri-decoder architecture into a single unified decoder by employing three specialized experts dedicated to nuclei segmentation, normal edge delineation, and cluster edge detection, respectively. This approach mirrors the functionality of tri-decoder in a more efficient and streamlined manner, achieving 32% reduction in parameters by sharing attention heads across the experts. Furthermore, MoE-NuSeg leverages the Transformer's [20] ability to capture global correlations to enhance segmentation performance.

Our key contributions are threefold. First, we introduce MoE-NuSeg, a novel domain-knowledge-driven MoE network for nuclei segmentation, built on the Swin Transformer architecture [19]. Second, we propose an innovative attention-based gating network that dynamically modulates the contributions of the three specialized experts in MoE-NuSeg based on the input data, enhancing segmentation performance and robustness. Finally, we design a novel two-stage training scheme. In the first stage, the three specialized experts are trained independently to excel in their respective tasks. This approach consolidates the roles previously distributed across three decoders into a single decoder, reducing the parameter count through shared attention heads. In the second stage, these experts are trained alongside the attention based gating network, fostering collaboration and co-evolution among the components. This two-stage training strategy allows the experts to leverage their specialized knowledge while dynamically collaborating based on the input data, ultimately leading to improved segmentation performance.

2. Related work

2.1. Evolution of single decoder architectures

U-Net [21] represents a pivotal advancement in automating medical image segmentation, marking the inception of the CNN era in this field. Its influential encoder–decoder architecture, featuring skip connections, remains widely used in contemporary applications. Subsequently, numerous variations [6,8,22,23] of U-Net gained attention for their simplicity and effectiveness.

The advent of transformers [20] has catalyzed a paradigm shift in architecture design, particularly with the emergence of vision transformers [24]. These attention-based architectures move beyond the local receptive fields of CNNs, enabling the capture of global spatial relationships in image processing [15,25]. The integration of transformers [20] into architectures such as TransUNet [26] represents a fusion of transformer capabilities with the U-Net architecture. The Swin Transformer [19] has notably demonstrated its efficacy in processing large-scale image datasets, which has led to the development of Swin-Unet [27] for segmentation purposes.

2.2. Foundation model

Recently, foundation models have gained prominence as powerful learning architectures, pre-trained on extensive datasets to capture a wide range of patterns and features. The Segment Anything Model (SAM) [28] demonstrates impressive zero-shot generalization capabilities in natural images. However, when applied to medical images, SAM experiences a notable performance drop due to the data distribution gap between natural and medical images. Several studies [29–33] have investigated adapting SAM for medical imaging by fine-tuning it with specific medical datasets, yielding promising results. Nevertheless, most of these efforts focus on images of organs, which differ visually from nuclei, presenting challenges similar to those encountered in the transition from natural to medical images. This highlights the need for specialized approaches for nuclei segmentation.

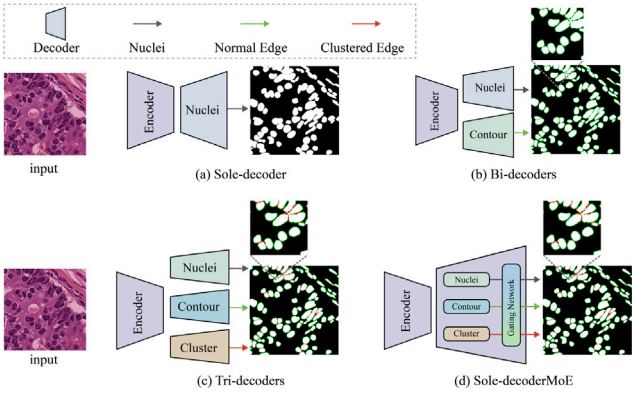

Fig. 1. (a) Single-decoder: A sole decoder performs nuclei segmentation [12]. (b) Bi-decoders: Two separate decoders handle nuclei and edge segmentation, respectively, enabling more precise delineation of nuclear boundaries [12,13]. (c) Tri-decoders: An additional decoder identifies clustered edges, further improving the separation of clustered nuclei [15]. (d) MoE-NuSeg: The tri-decoder architecture is replaced by a MoE structure, in which specialized experts (nuclei, edge, and cluster edge) share attention heads, and their contributions are dynamically combined through a gating network.

2.3. Advances in multi-task learning approaches

Due to the frequently clustering of nuclei, single-decoder architectures such as U-Net and its variants often struggle in scenarios where edge information is insufficient. The Deep Contour-Aware Neural Network (DCAN) [12] addresses this issue by employing a bi-decoder structure within a multi-task learning framework: one decoder focuses on nuclei segmentation, while the other extracts contours to identify nuclei edges. Similarly, the Contour-aware Informative Aggregation Network (CIA-Net) [13] enhances segmentation performance by integrating a multi-level information aggregation module between two task-specific decoders. This approach bidirectionally aggregates features tailored to each task, capturing spatial and textural dependencies between nuclei and contours, thereby improving edge delineation and spatial relationships in segmentation tasks. More recently, $CA^{2.5}$-Net [14] has demonstrated notable performance by emphasizing cluster edge detection. Additionally, TransNuSeg [15] introduced three decoder architectures within a multitask learning framework, leading to notable improvements in nuclei segmentation performance.

3. Methodology

3.1. Overview of MoE-NuSeg

The overall architecture of MoE-NuSeg, as illustrated in Fig. 2, follows a classic U-shaped encoder–decoder design. The hierarchical structure of the encoder facilitates feature extraction across multiple levels of granularity, from low to high. MoE-NuSeg incorporates dual Swin-Transformer-based blocks [19] across all three layers of the encoder's down-sampling pathways, effectively capturing both local and long-range feature correlations [15]. In the decoder, we employ novel

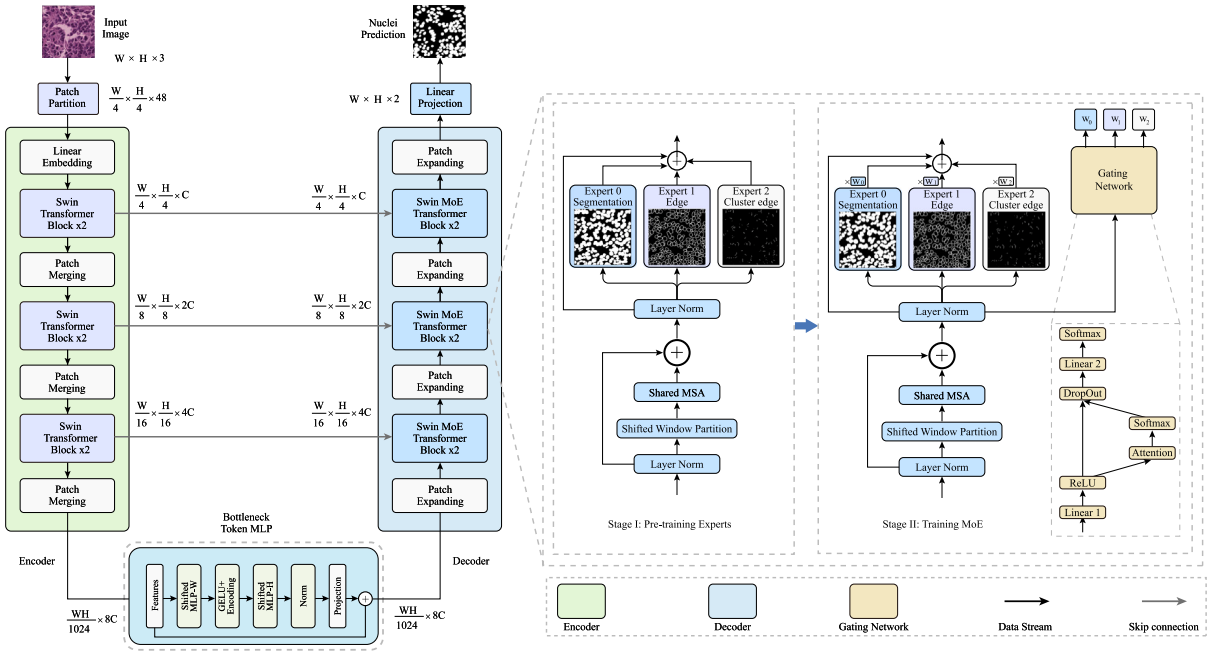

Fig. 2. Architecture of MoE-NuSeg. The encoder consists of Swin Transformer blocks that extract multi-scale features, while the decoder employs MoE-based Swin Transformer blocks for task-specific segmentation of nuclei, normal edges, and clustered edges. Skip connections link the encoder to the decoder, providing the detailed feature signals for reconstruction. The gating network dynamically allocates expert contributions, with task-specific output heads generating the final predictions. The diagram illustrates both Stage I (Pre-training experts) and Stage II (Training MoE).

MoE-based Swin Transformer blocks across its three layers, strategically decoupling the complex nuclei segmentation challenge into three interrelated, domain-knowledge-driven subtasks: nuclei segmentation, normal edge segmentation, and cluster edge segmentation. Leveraging the proposed two-stage learning paradigm, MoE-NuSeg surpasses multi-decoder architectures by enabling more efficient parameter sharing and reducing redundant feature learning. The encoder and decoder are connected through a bottleneck token MLP [15,34], offering a parameter-efficient alternative to the traditional Swin Transformer bottleneck. To preserve and recover image details, skip connections are utilized to link all three Swin Transformer-based blocks in the encoder with their corresponding layers in the decoder, a strategy widely adopted in segmentation models [5].

MoE-NuSeg adopts a multi-task learning framework, prioritizing nuclei segmentation as the primary task while treating the segmentation of normal edges and clustered edges as auxiliary tasks. By incorporating additional supervision signals from both normal and clustered edges through their respective experts, the model aims to enhance overall performance and improve its ability to accurately delineate nuclei boundaries and separate adjacent nuclei. As a result, the training objective is designed to optimize performance across all three decoder branches using a combined loss function, which integrates both cross-entropy loss and Dice loss. By optimizing this multi-task learning objective, MoE-NuSeg effectively leverages the specialized knowledge of each expert while addressing the interdependencies among the subtasks of nuclei segmentation, normal edge segmentation, and clustered edge segmentation.

3.2. Domain-knowledge-based mixture of experts

To improve nuclei segmentation performance, we decompose the task into three specialized subtasks: nuclei segmentation, normal edge segmentation, and clustered edge segmentation, utilizing both nuclei edge and clustered edge information. Each subtask is assigned to an expert within the Swin Transformer block. Formally, the architecture of each expert is defined as follows:

$$\texttt{Expert}(x) = \texttt{Dropout}(\texttt{Linear}(\texttt{Dropout}(\texttt{GELU}(\texttt{Linear}(x))))) \quad (1)$$

where x represents the input features to the expert. Each expert consists of two linear layers, separated by the GELU activation function and dropout layers. The first linear layer projects the input features to a higher-dimensional space, followed by a GELU activation function that introduces nonlinearity. The dropout layers are used to mitigate overfitting by randomly setting a fraction of the activations to zero during training. The second linear layer projects the transformed features back to the original dimension, and the final dropout layer helps further regulate the output. Each expert serves a unique purpose in the nuclei segmentation task:

- **Nuclei Segmentation Expert:** This expert classifies each pixel as either foreground (nucleus) or background, producing a binary segmentation map that identifies regions occupied by nuclei.
- **Normal Edge Expert:** Specializing in boundary detection, this expert outlines the contours of each nucleus, generating a binary map that clearly defines nuclei boundaries and distinguishes them from the background.
- **Cluster Edge Expert:** Focused on densely packed nuclei, this expert detects cluster edges, creating a binary map that emphasizes the contiguous edges of adjacent nuclei for clearer segmentation. Fig. 3 shows the attention map of the cluster edge expert in layer one of the decoder.

By training each expert on a specific subtask, the experts learn specialized features that enhance the overall segmentation performance of MoE-NuSeg in the second stage. As a result, our MoE method streamlines the segmentation process by leveraging the specialized knowledge of each expert within a single Swin Transformer block, avoiding the need for multiple decoders and reducing the number of parameters.

3.3. Two-stage training scheme

MoE-NuSeg employs a two-stage training scheme to simultaneously optimize three experts within the MoE framework. In the first stage, each expert is trained independently to specialize in a specific subtask: nuclei segmentation, normal edge segmentation, and cluster edge segmentation. This stage emphasizes developing task-specific expertise.

In the second stage, an attention-based gating network is introduced to guide the collaboration among the three experts. The gating network and the experts are then trained jointly, fostering co-evolution and enhancing collaboration. Our two-stage training scheme not only strengthens the collaborative dynamics between the experts but also improves the overall efficacy of the segmentation process.

Algorithm 1: Two-Stage Training of MoE-NuSeg

Stage I: Training of Three Specialized Experts
Input: Training data: X; Ground truth: Y
Output: Trained nuclei segmentation, normal edge, and cluster edge experts

Initialize Experts: $\text{Expert}_{\text{nuclei}}$, $\text{Expert}_{\text{edge}}$, $\text{Expert}_{\text{cluster}}$
for $epoch = 1$ to M **do**
  $\widehat{Y_{\text{n}}}, \widehat{Y_{\text{e}}}, \widehat{Y_{\text{c}}} \leftarrow \text{model}(X)$
  Compute $\mathcal{L}_n$, $\mathcal{L}_e$, $\mathcal{L}_c$ as defined in Eq. (3)
  Compute $\mathcal{L}_{total}$ as defined in Eq. (2)
  Compute $\theta^* = \arg\min_\theta \mathcal{L}_{total}(\theta)$
end for

Stage II: Training for Gating Network Integration
Input: Training data X; Ground truth Y, Trained model parameters θ from Stage I
Output: Fully trained MoE-NuSeg, including gating network

Initialize Gating Network
for $epoch = 1$ to M **do**
  $w_i \leftarrow \text{GatingNetwork}(X)$
  $w_i \leftarrow \text{SoftMax}(\gamma \cdot w_{\text{pre}} + (1-\gamma) \cdot w_i)$
  $\hat{y}_i \leftarrow \sum_{i=1}^{3} w_i' \cdot \texttt{Expert}_i(x)$
  Compute $\mathcal{L}_{total}$ as defined in Eq. (2)
  Compute $\theta^* = \arg\min_\theta \mathcal{L}_{total}(\theta)$
end for

3.3.1. Stage I: Specialized training of three experts

The first training stage focuses on developing the expertise of the three experts independently. Each expert is trained to tackle a distinct subtask of nuclei segmentation: nuclei segmentation, normal edge segmentation, and cluster edge segmentation. The training process optimizes model parameters by minimizing the total loss $\mathcal{L}_{total}$, which is weighted sum of three individual loss components of each expert: nuclei loss $\mathcal{L}_n$, edge loss $\mathcal{L}_e$ and cluster edge loss $\mathcal{L}_c$, mathematically, it is formulated as:

$$\mathcal{L}_{total} = \gamma_n \cdot \mathcal{L}_n + \gamma_e \cdot \mathcal{L}_e + \gamma_c \cdot \mathcal{L}_c \tag{2}$$

The weights assigned to each loss component are represented by the coefficients γ_n, γ_e, and γ_c, which are set to 0.30, 0.35, and 0.35, respectively [15].

To compute nuclei loss $\mathcal{L}_n$, edge loss $\mathcal{L}_e$, and cluster edge loss $\mathcal{L}_c$, we adhere to a weighting scheme of 0.40 for cross-entropy loss and 0.60 for Dice similarity coefficient (DSC) loss, as specified in prior work [14]. This weighting ensures that the model effectively learns to classify pixels accurately (via Cross-Entropy loss) while promoting a overlap between predicted and ground-truth segmentation masks (via Dice loss).

The nuclei loss $\mathcal{L}_n$ is defined as follows:

$$\mathcal{L}_n = 0.6 \cdot \mathcal{L}_{Dice}(y_n, \hat{y_n}) + 0.4 \cdot \mathcal{L}_{CE}(y_n, \hat{y_n}) \tag{3}$$

where y_n and $\hat{y_n}$ denote the ground truth and predicted values for nuclei segmentation, respectively. The edge loss $\mathcal{L}_e$ and cluster edge loss $\mathcal{L}_c$ are formulated similarly.

During this stage, each expert is trained on its respective ground truth, tailored to its specific subtask. The experts independently learn to predict nuclei instances, normal edge masks, and clustered edge masks with accuracy. This stage ensures that each expert becomes proficient in capturing essential details relevant to its designated subtask, similar to tri-decoder nuclei segmentation models [15]. Training in this stage allows the experts to acquire specialized representations without interference from the gating network, enhancing their contributions in the subsequent collaborative stage.

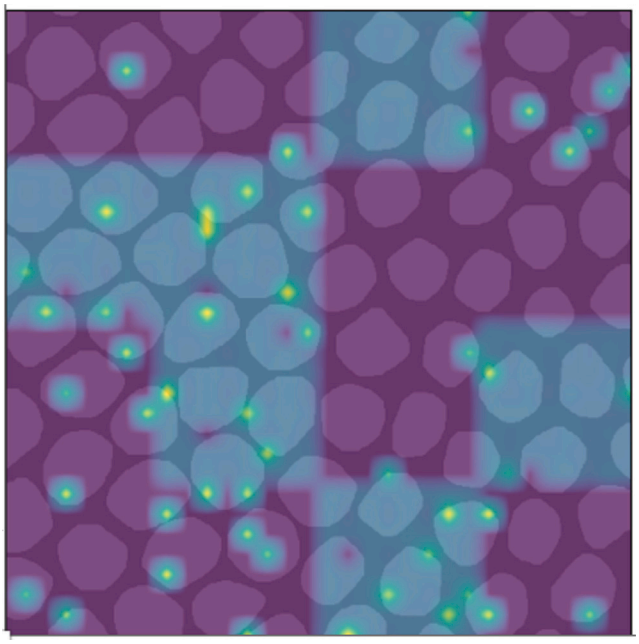

Fig. 3. The attention map of the cluster edge expert, averaged across the heads of the first-layer transformer block, shows attention concentrated along nuclei boundaries, represented by green blobs. This indicates the model's focus on critical regions for improving segmentation accuracy. The purple grid represents Swin Transformer windows, which divide the image into smaller blocks, allowing the model to process distinct regions using shifted windows.

3.3.2. Stage II: Gating network integration training

In the second training stage, a gating network is introduced to effectively integrate the outputs from the three experts in each layer of decoder, improving overall segmentation performance. The output from the preceding Multi-head Self-Attention (MSA) is passed to corresponding trained experts. The gating network then learns to dynamically allocate each expert's contributions based on the input x. The combined outputs of the experts are merged using a weighted sum:

$$x_{out} = \sum_{i=1}^{3} w_i' \cdot \texttt{Expert}_i(x_{in}) \tag{4}$$

where x_{in} and x_{out} represent the input and output of the MoE, respectively. The weights w_i' ($i \in \{0, 1, 2\}$ denoting the index of the expert) are calculated by applying the softmax function to a linear combination of the output of the gating network w_i and the predefined weights w_{pre}. The predefined weights w_{pre} are based on data type, specifically nuclei, the edges, and the cluster edge masks. A mixing coefficient $\gamma \in (0, 1)$ is introduced to modulate the influence of w_{pre} on the final weight calculation, finally, a softmax function is applied to the combined weights to derive the final weights for the three experts:

$$w_i' = \texttt{SoftMax}(\gamma \cdot w_{pre} + (1-\gamma) \cdot w_i) \tag{5}$$

This formulation integrates the rule-based prior knowledge stored in w_{pre} with the dynamic insights w_i generated by the attention-based gating network. It allows for a flexible combination of experts' outputs, ensuring that the integration is both responsive to the data and grounded in established knowledge. Predefined weights w_{pre} are assigned based on the expert type and the input data specific to the experts. For instance, the output from the cluster edge expert is assigned a

Table 1
Performance comparison of MoE-NuSeg with the state-of-the-art nuclei segmentation models on three diverse datasets. The evaluation metrics include DSC, F1 score, IoU, Accuracy, and Precision. The results are presented as mean value with the 95% confidence interval ($n = 6$), with the highest performance for each metric, emphasized in **boldface**. MoE-NuSeg consistently outperforms comparable models across most metrics and datasets, demonstrating its robustness and effectiveness in nuclei segmentation tasks. For the one-shot generalization evaluation of SAM and SAM-Med2D, the models are subjected to a single inference process without fine-tuning on the nuclei datasets. Consequently, the confidence intervals are not calculated for this assessment.

Dataset	Metrics	SegNet	U-Net	TransUNet	Swin-Unet	CA$^{2.5}$-Net	TransNuSeg	SAM	SAM-Med2D	MoE-NuSeg (ours)
Histology	DSC	75.08$_{\pm 2.84}$	83.18$_{\pm 0.39}$	79.10$_{\pm 1.60}$	77.33$_{\pm 0.65}$	81.90$_{\pm 1.20}$	83.21$_{\pm 0.60}$	68.01	73.50	**85.13**$_{\pm 0.42}$
	F1	82.87$_{\pm 6.66}$	83.27$_{\pm 2.49}$	79.59$_{\pm 7.70}$	76.94$_{\pm 5.28}$	**85.37**$_{\pm 6.92}$	83.34$_{\pm 3.59}$	51.97	63.69	84.71$_{\pm 3.07}$
	IoU	62.94$_{\pm 2.45}$	71.30$_{\pm 0.55}$	65.89$_{\pm 1.97}$	63.22$_{\pm 0.88}$	71.09$_{\pm 1.33}$	71.38$_{\pm 0.86}$	54.13	60.74	**73.19**$_{\pm 1.02}$
	Accuracy	87.27$_{\pm 0.48}$	90.22$_{\pm 0.22}$	88.42$_{\pm 0.44}$	88.11$_{\pm 0.33}$	90.45$_{\pm 0.59}$	90.33$_{\pm 0.34}$	72.48	79.80	**91.38**$_{\pm 0.35}$
	Precision	80.55$_{\pm 1.65}$	82.34$_{\pm 0.70}$	82.23$_{\pm 1.33}$	**85.03**$_{\pm 0.61}$	83.88$_{\pm 1.31}$	83.31$_{\pm 0.83}$	62.32	65.82	84.82$_{\pm 1.00}$
Fluorescence	DSC	93.35$_{\pm 1.42}$	95.50$_{\pm 0.50}$	94.62$_{\pm 0.61}$	93.83$_{\pm 0.65}$	93.89$_{\pm 0.75}$	95.78$_{\pm 0.48}$	80.65	81.36	**95.93**$_{\pm 0.39}$
	F1	94.50$_{\pm 3.64}$	96.52$_{\pm 1.03}$	94.85$_{\pm 2.39}$	95.64$_{\pm 3.61}$	95.68$_{\pm 3.47}$	96.44$_{\pm 1.57}$	83.51	84.27	**96.67**$_{\pm 1.48}$
	IoU	87.23$_{\pm 1.47}$	91.56$_{\pm 0.89}$	89.92$_{\pm 1.06}$	88.67$_{\pm 1.09}$	88.81$_{\pm 1.24}$	91.84$_{\pm 0.84}$	71.71	72.22	**92.26**$_{\pm 0.67}$
	Accuracy	94.66$_{\pm 0.70}$	96.40$_{\pm 0.32}$	95.56$_{\pm 0.56}$	95.43$_{\pm 0.40}$	95.28$_{\pm 0.47}$	96.54$_{\pm 0.37}$	86.93	90.35	**96.73**$_{\pm 0.38}$
	Precision	95.44$_{\pm 1.32}$	95.59$_{\pm 0.68}$	94.24$_{\pm 1.01}$	93.28$_{\pm 1.32}$	93.61$_{\pm 1.53}$	95.44$_{\pm 0.85}$	80.26	84.19	**95.61**$_{\pm 0.65}$
Lizard	DSC	72.68$_{\pm 2.02}$	72.41$_{\pm 1.47}$	64.12$_{\pm 3.57}$	61.92$_{\pm 3.25}$	64.80$_{\pm 4.31}$	76.46$_{\pm 1.15}$	6.32	6.05	**77.36**$_{\pm 1.12}$
	F1	73.18$_{\pm 3.24}$	72.96$_{\pm 2.19}$	65.64$_{\pm 2.51}$	65.54$_{\pm 5.01}$	72.41$_{\pm 6.69}$	75.94$_{\pm 2.27}$	6.05	6.32	**77.09**$_{\pm 2.54}$
	IoU	57.30$_{\pm 2.30}$	56.88$_{\pm 1.75}$	47.85$_{\pm 3.46}$	45.17$_{\pm 3.44}$	49.90$_{\pm 3.81}$	61.98$_{\pm 1.49}$	3.34	3.18	**63.16**$_{\pm 1.48}$
	Accuracy	88.57$_{\pm 0.87}$	88.47$_{\pm 0.77}$	85.87$_{\pm 1.32}$	87.00$_{\pm 0.52}$	86.33$_{\pm 0.75}$	90.02$_{\pm 0.67}$	82.53	82.60	**90.44**$_{\pm 0.59}$
	Precision	72.14$_{\pm 0.94}$	72.03$_{\pm 1.04}$	68.09$_{\pm 1.99}$	75.61$_{\pm 0.85}$	**77.90**$_{\pm 2.96}$	75.18$_{\pm 0.96}$	36.57	35.78	76.29$_{\pm 0.90}$

predefined weight of 1 for its specific task, while the other two experts are assigned a weight of 0. The outputs generated by the attention-based gating network, w_i $(i \in 0, 1, 2)$, serve as dynamic weights that adjust in response to the variability of the input data. This process mirrors how the human brain uses heuristics to efficiently process information and make decisions. The pre-determined weights in the model act as cognitive shortcuts, simplifying decision-making [35,36]. These weights provide an initial baseline, allowing the model to learn from data and optimize its performance over time.

3.4. Attention-based gating network

The aggregation of expert outputs within MoE-NuSeg is managed by a novel attention-based gating network, which is trained to evaluate and assimilate each expert's contributions. Inspired by the dynamic flexibility of Transformers [20] and principles from cognitive science [37], the gating network's outputs are influenced by the input data, ensuring that the fusion of expert outputs is both contextually aware and tailored to the specific data type presented. At the core of this design is an MLP, augmented with an attention mechanism, that intelligently responds to input data, enabling flexible and adaptive expert integration. The attention mechanism in the gating network is implemented as follows:

$$\alpha = \texttt{SoftMax}(\texttt{Attention}(\texttt{ReLU}(\texttt{Linear}(x)))) \tag{6}$$

where input features x are first linearly transformed and passed through a ReLU activation function. An attention function is then applied to compute attention scores, which are normalized using the Softmax function to obtain the attention weights α. The gating network itself is defined as:

$$\texttt{Gating}(x) = \texttt{Linear}(\texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(x) \cdot \alpha))) \tag{7}$$

The input features x are linearly transformed and element-wise multiplied by the attention weights α. The result is then passed through a ReLU activation function and a dropout layer for regularization. Finally, another linear transformation is applied to obtain the gating network's output. The attention-based gating network in MoE-NuSeg, inspired by cognitive science, enhances decision-making through adaptive data interpretation. This approach is grounded in an attention mechanism that dynamically adjusts weights in real-time based on the input data, allowing it to capture long-range global dependencies. This strategy has proven effective in Transformer architectures [20], which outperform CNNs that rely on fixed weights and exhibit an inductive bias toward local features.

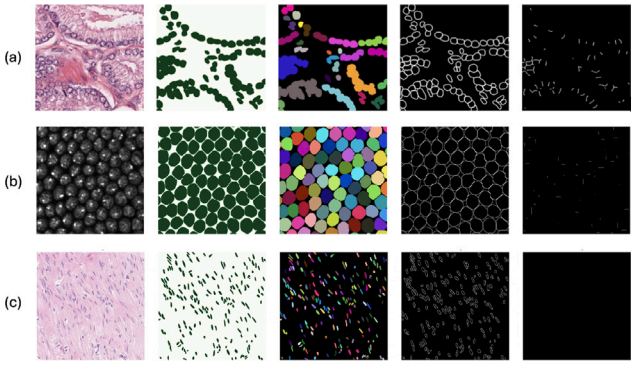

Fig. 4. Visualization of the annotation process in MoE-NuSeg. The first column displays raw images from three distinct datasets: (a) fluorescence microscopy, (b) histology, and (c) lizard. The second and third columns present their corresponding semantic and instance masks. Columns four and five show the normal edge segmentation and clustered edge segmentation, respectively. These derived masks serve as ground truth for training the corresponding experts, which are integrated into each layer of the decoder.

4. Experiments

4.1. Datasets

We evaluate the models' performance across three distinct datasets, primarily focusing on histology images, complemented by fluorescence microscopy images to demonstrate the adaptability and robustness of MoE-NuSeg beyond histology applications. The core datasets used in our experiments are as follows:

- **Histology Image Dataset**[1]: This dataset combines the publicly available MoNuSeg dataset [38] with a in-house histology dataset [11], resulting in 462 histology images. The MoNuSeg dataset was preprocessed by dividing each image into four partially overlapping patches, each with a resolution of 512 × 512 pixels. Additionally, the proprietary dataset comprises 300 images extracted from 50 whole-slide images (WSIs) scanned at 20× magnification. All in-house images were annotated by five pathologists according to MoNuSeg guidelines, ensuring high-quality ground truth for histology-based evaluations.

[1] https://github.com/lu-yizhou/ClusterSeg.

- **Lizard Image Dataset** [39]: This dataset, consisting of nearly half a million labeled nuclei in H&E-stained colon tissue, is recognized as the largest dataset for nuclei instance segmentation to date. It stands out for its pathologist-in-the-loop refinement process, ensuring high-quality annotations. Since the original download link from the publication is no longer active, we obtained the dataset from an alternative repository on Kaggle,[2] which includes a subset of 238 images with a resolution of 224 × 224 pixels, along with their corresponding labels.
- **Fluorescence Microscopy Image Dataset** [40]: This dataset comprises three heterogeneous data sources,[3] totaling 524 fluorescence images, each with a resolution of 512 × 512 pixels. The diversity of this dataset provides a rigorous benchmark for evaluating MoE-NuSeg's performance on microscopy images with varying characteristics.

To further validate the versatility of MoE-NuSeg, we expanded our evaluation to include two additional well-known datasets, TNBC [41] and CPM-17 [42], which feature diverse nuclei shapes and tissue types, allowing us to assess the model's robustness across different imaging modalities and clinical scenarios. Detailed discussions of these datasets are provided in Section 4.5.

To standardize the data for training and evaluation, we performed normalization preprocessing using OpenCV (CV2) [43,44]. For datasets without instance segmentation, CV2 was used to generate the required masks. Furthermore, all image files were resized to a uniform resolution of 512 × 512 pixels to ensure consistency across all datasets. Fig. 4 illustrates representative samples from the three datasets, accompanied by their corresponding processed masks.

4.2. Evaluation metrics

To comprehensively evaluate the performance of MoE-NuSeg, we employ the following metrics in alignment with previous work [45]:

- **Dice Similarity Coefficient (DSC):** Measures the overlap between the predicted segmentation and the ground truth, emphasizing true positives.
- **Intersection over Union (IoU):** Assesses the ratio of the intersection to the union of predicted and ground truth segmentation.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced assessment of model performance.
- **Accuracy:** The ratio of correct predictions to the total number of predictions.
- **Precision:** The proportion of correctly predicted positive cases among all positive predictions.

4.3. Implementation details

To ensure fair comparisons, we implemented a uniform data split strategy across all datasets. Each dataset was randomly divided into three subsets: 80% for training, 10% for validation, and 10% for testing. During training, we used the Adam optimizer with an initial learning rate of 1×10^{-3}. Early stopping was implemented if the validation loss did not improve for 50 epochs. Each expert was modeled as a 2-layer MLP, with the output dimension matching the input, and the hidden layer dimension was set to four times the input size, using the GELU activation function. The number of attention heads in each of the three decoder layers increases sequentially: 6, 12, and 24 for the first through third layers, respectively.

[2] https://www.kaggle.com/datasets/aadimator/lizard-dataset/data.
[3] https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BSST265.

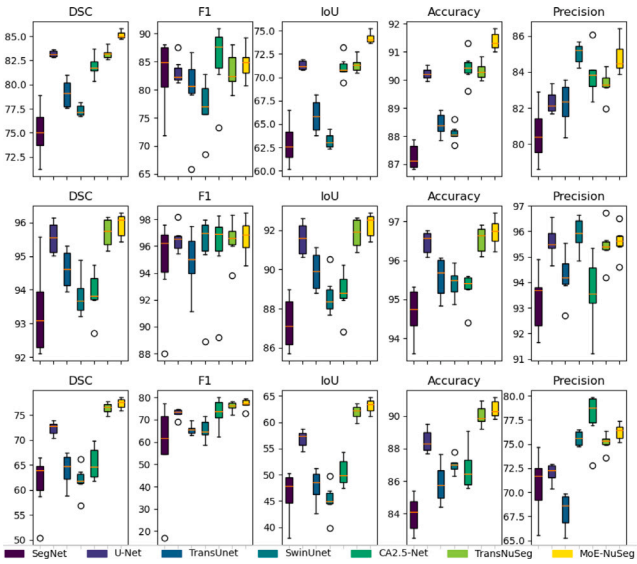

Fig. 5. Boxplot comparison of metrics across three datasets, arranged from top to bottom: histology, fluorescence, and lizard. MoE-NuSeg consistently demonstrates improved performance relative to competing methods, exhibiting reduced performance variability, which underscores its robustness and reliability in diverse clinical scenarios.

To promote effective learning and ensure robust convergence, we adopted the Cosine Annealing learning rate schedule, which gradually adjusts the learning rate. This strategy enables a more nuanced exploration of the optimization landscape. The maximum number of iterations (T_max) was set to 2000, with the minimum learning rate value (η_min) set to 1×10^{-8}.

All experiments were conducted on a single NVIDIA RTX 4090 GPU with 24 GB of VRAM, paired with a 16 vCPU Intel(R) Xeon(R) Gold 6430 processor and 60 GB of system RAM.

4.4. Results

Table 1 provides a comprehensive comparison of MoE-NuSeg with state-of-the-art models, including SegNet, U-Net, TransUNet, Swin-Unet, CA$^{2.5}$-Net, and TransNuSeg. The results are reported as the mean and 95% confidence interval (CI) across six runs. On the Histology dataset, MoE-NuSeg achieved the highest performance in terms of DSC (85.13%), F1 score (84.71%), IoU (73.19%), and Accuracy (91.38%). While Swin-Unet attained the highest Precision (85.03%), MoE-NuSeg closely followed with a Precision of 84.82%. These results highlight MoE-NuSeg's enhanced ability to accurately segment nuclei in histological images compared to the other models. For the Fluorescence dataset, MoE-NuSeg outperformed all models across all metrics, achieving a DSC of 95.93%, F1 score of 96.67%, IoU of 92.26%, Accuracy of 96.73%, and Precision of 95.61%. These results underscore MoE-NuSeg's robustness and effectiveness in handling fluorescence images, which often present different challenges compared to histological images. On the Lizard dataset, MoE-NuSeg also demonstrated strong performance, achieving a DSC of 77.36%, F1 score of 77.09%, IoU of 63.16%, and Accuracy 90.44%. While CA$^{2.5}$-Net achieved the highest Precision (77.90%), slightly surpassing MoE-NuSeg's Precision of 76.29%, MoE-NuSeg remained highly competitive, showcasing its adaptability across diverse image domains. As shown in Fig. 5, MoE-NuSeg consistently outperformed the other models across most metrics with minimal performance variation, demonstrating its robustness and effectiveness in nuclei segmentation. The integration of domain-knowledge-driven Mixture of Experts (MoEs) and the attention-based gating network enables MoE-NuSeg to efficiently capture and incorporate specialized

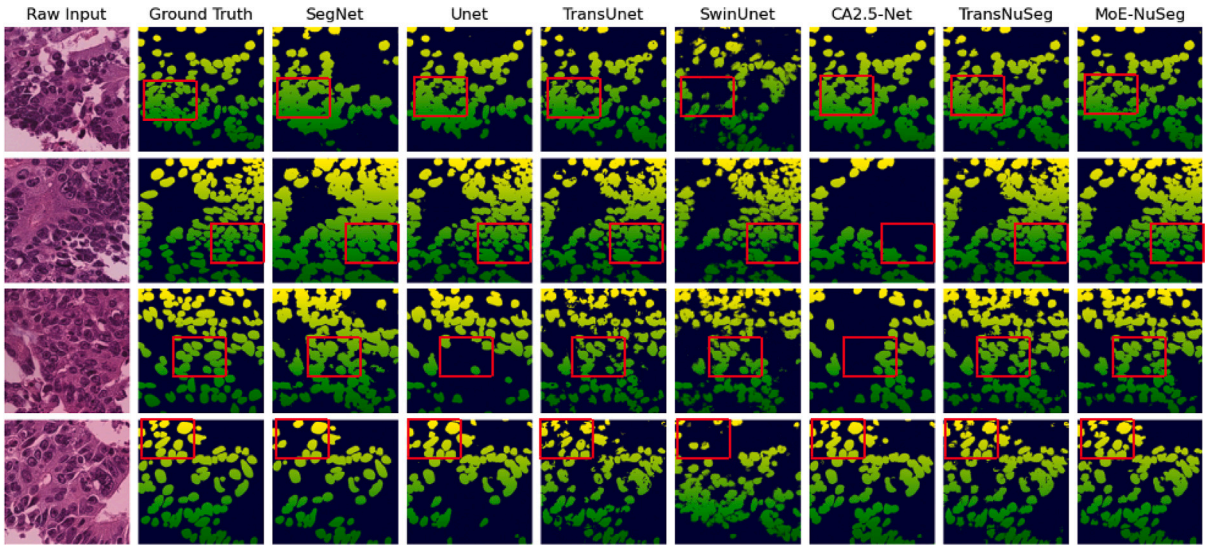

Fig. 6. Visual comparison of nuclei segmentation results across different models on representative images. Areas highlighted in red boxes showcase regions where MoE-NuSeg excels, particularly in accurately delineating nuclei boundaries, effectively separating clustered nuclei, and adeptly handling challenging cases with irregularly shaped nuclei.

Table 2
Paired one-tailed t-test p-values comparing the performance of MoE-NuSeg (M) with other methods (O) on three datasets. The null hypothesis $H_0 : \mu_M \leq \mu_O$ suggests MoE-NuSeg does **not** outperform other methods, while the alternative hypothesis $H_1 : \mu_M > \mu_O$ suggests it does. Consistently low p-values strongly indicate MoE-NuSeg's enhanced performance relative to the compared methods.

Methods	Metrics	Histology	Fluorescence	Lizard
SegNet	DSC	9.26×10^{-5}	1.26×10^{-3}	5.99×10^{-4}
	F1	2.30×10^{-1}	1.63×10^{-1}	4.29×10^{-2}
	IoU	2.27×10^{-5}	1.94×10^{-5}	1.43×10^{-4}
	Acc	2.90×10^{-6}	2.14×10^{-5}	4.51×10^{-6}
	Precision	2.91×10^{-3}	2.61×10^{-3}	1.80×10^{-3}
U-Net	DSC	2.37×10^{-5}	2.09×10^{-2}	8.43×10^{-6}
	F1	2.75×10^{-2}	3.05×10^{-1}	1.82×10^{-5}
	IoU	2.73×10^{-5}	1.90×10^{-2}	6.43×10^{-6}
	Acc	3.93×10^{-6}	2.38×10^{-2}	2.45×10^{-6}
	Precision	1.30×10^{-3}	4.97×10^{-1}	6.00×10^{-5}
TransUNet	DSC	4.12×10^{-5}	1.64×10^{-4}	3.10×10^{-5}
	F1	2.67×10^{-2}	3.93×10^{-2}	4.22×10^{-5}
	IoU	2.15×10^{-5}	1.48×10^{-4}	9.06×10^{-6}
	Acc	1.12×10^{-5}	1.57×10^{-3}	1.41×10^{-5}
	Precision	2.71×10^{-4}	3.27×10^{-3}	1.22×10^{-4}
Swin-Unet	DSC	3.32×10^{-6}	1.08×10^{-4}	4.95×10^{-5}
	F1	6.56×10^{-3}	3.01×10^{-1}	1.37×10^{-3}
	IoU	3.25×10^{-6}	9.40×10^{-5}	3.50×10^{-5}
	Acc	1.21×10^{-5}	1.33×10^{-3}	8.29×10^{-6}
	Precision	3.33×10^{-1}	1.63×10^{-1}	2.33×10^{-2}
$CA^{2.5}$-Net	DSC	2.88×10^{-4}	1.04×10^{-4}	7.36×10^{-5}
	F1	4.05×10^{-1}	3.03×10^{-1}	1.41×10^{-1}
	IoU	7.46×10^{-4}	7.56×10^{-5}	2.23×10^{-5}
	Acc	1.21×10^{-2}	2.32×10^{-4}	3.45×10^{-4}
	Precision	1.05×10^{-1}	4.99×10^{-3}	1.18×10^{-1}
TransNuSeg	DSC	9.35×10^{-4}	2.31×10^{-2}	5.74×10^{-4}
	F1	4.42×10^{-2}	9.14×10^{-2}	7.97×10^{-4}
	IoU	9.41×10^{-4}	2.15×10^{-2}	4.94×10^{-4}
	Acc	9.79×10^{-4}	1.52×10^{-2}	2.02×10^{-3}
	Precision	2.52×10^{-3}	6.46×10^{-2}	4.03×10^{-2}

knowledge, leading to improved segmentation accuracy and adaptability across different image modalities. An example of segmentation results from different models is illustrated in Fig. 6.

To further validate the statistical significance of MoE-NuSeg's performance, we conducted paired one-tailed t-tests [46,47], comparing MoE-NuSeg (denoted as M) with each of the other methods (denoted as O) across all datasets and evaluation metrics, using the same random seeds (n = 6). The null hypothesis $H_0 : \mu_M \leq \mu_O$ posits that MoE-NuSeg **does not outperform** the compared method, while the alternative hypothesis $H_1 : \mu_M > \mu_O$ asserts that MoE-NuSeg **performs better**.

Table 2 presents the p-values from the paired one-tailed t-tests. Consistently low p-values across all datasets and metrics provide strong evidence against the null hypothesis, supporting the effectiveness of MoE-NuSeg over the other methods. For example, when comparing MoE-NuSeg with TransNuSeg on the Histology dataset, the p-values for DSC, F1, IoU, Accuracy, and Precision are 9.35×10^{-4}, 4.42×10^{-2}, 9.41×10^{-4}, 9.79×10^{-4}, and 2.52×10^{-3}, respectively, all well below the commonly accepted significance level of 0.05. Similar patterns of low p-values were observed for the comparisons with other methods across the Fluorescence and Lizard datasets, further reinforcing the statistical significance of MoE-NuSeg's improved performance.

4.5. Further evaluation: TNBC & CPM-17

To further validate the versatility of the model, we expanded the evaluation to two additional widely recognized datasets: TNBC [48] and CPM-17 [42], using a random seed of 41. TNBC presents a challenging array of diverse cell types, making it ideal for testing the model's ability to generalize across different cellular morphologies. Meanwhile, CPM-17 introduces complexity with its variety of nuclei shapes and densely packed cells, providing a test of the model's robustness. Both datasets are clinically relevant and serve as benchmarks in medical imaging.

The TNBC dataset comprises 50 images with 4022 annotated cells, encompassing cell types such as normal epithelial cells, invasive cancer cells, and immune cells. Each sample includes between 5 and 293 cells, with an average of 80 cells per sample. Annotations were performed by three experts – a pathologist and two research fellows – who resolved disagreements through consensus. Conversely, the CPM-17 dataset includes 64 images of various brain cancer types, featuring 7570 annotated nuclei. These images, scanned at 20x and 40x magnifications, provide detailed insights into nuclei across cancers such as non-small cell lung cancer (NSCLC), head and neck squamous cell carcinoma (HNSCC), glioblastoma multiforme (GBM), and lower-grade glioma (LGG), posing significant challenges to the model with complex cancerous nuclei.

The results presented in Table 3 underscore the performance advantages of MoE-NuSeg over the state-of-the-art TransNuSeg model. This comparison demonstrates MoE-NuSeg's enhanced ability to generalize across various imaging modalities and tissue types, further reinforcing its versatility.

Table 3
Performance comparison of MoE-NuSeg and TransNuSeg across the TNBC and CPM-17 datasets.

Dataset	Metric	MoE-NuSeg (%)	TransNuSeg (%)
TNBC	DSC	**78.97**	74.51
	F1	**70.67**	64.67
	IoU	**66.03**	60.36
	Accuracy	**93.66**	93.09
	Precision	78.82	**81.11**
CPM-17	DSC	**82.23**	82.01
	F1	**84.89**	78.13
	IoU	**69.90**	69.07
	Accuracy	**91.73**	91.09
	Precision	**81.77**	78.62

4.6. Model complexity and computation efficiency

In addition to segmentation performance, we assessed the complexity and computational efficiency of MoE-NuSeg against state-of-the-art models. Table 4 compares of the number of parameters, FLOPs (floating point operations), and training/inference times for SegNet, U-Net, TransUNet, Swin-Unet, CA$^{2.5}$-Net, TransNuSeg, and two variants of MoE-NuSeg: stage one only (I) and two stages (I+II). MoE-NuSeg (I) and MoE-NuSeg (I+II) have 20.39 million and 20.95 million parameters, respectively, lower than most compared models. Only U-Net (10.55 million) has fewer parameters, highlighting the compact design and ability to minimize memory requirements for storage and deployment. In terms of computational cost, MoE-NuSeg (I) requires 59.72 billion FLOPs, comparable to most models, except for U-Net (42.74 $\times 10^9$) and Swin-Unet (31.13 $\times 10^9$). The two-stage variant, MoE-NuSeg (I+II), with 116.01 billion FLOPs, remains more efficient than models like TransUNet (130.41 billion), SegNet (160.52 billion), TransNuSeg (193.93 billion), and CA$^{2.5}$-Net (460.29 billion). This balance between complexity and efficiency underscores the advantages of its architecture.

Regarding training and inference time per 100 images, MoE-NuSeg (I) requires 37.71 s for training and 1.07 s for inference, while MoE-NuSeg (I+II) takes 13.61 s for training and 2.36 s for inference. Although MoE-NuSeg (I) has a longer training time than most models, its competitive inference time makes it suitable for real-time applications. MoE-NuSeg (I+II) achieves a balanced trade-off between training time and inference speed. Notably, CA$^{2.5}$-Net has the shortest training and inference time (8.27 s and 0.68 s, respectively) among all compared models. However, it also has the highest computational cost in terms of FLOPs (460.29 billion), indicating a trade-off between speed and complexity.

Overall, the model complexity comparison underscores the effectiveness of MoE-NuSeg in balancing size, computational efficiency, and performance. The proposed architecture, coupled with the two-stage training scheme, enables MoE-NuSeg to achieve competitive segmentation accuracy while reducing model complexity and inference time, making it well-suited for practical nuclei segmentation tasks.

4.7. Ablation study

We first conducted an ablation study on hyperparameter γ, which modulates the balance between prior weights (w_{pre}) and data driven weights (w_i) in the gating mechanism, as described in Eq. (5). To assess the effect of γ on the model's performance and determine the optimal value, we used a grid search method, varying γ from 0 to 1 in increments of 0.25. Fig. 7 presents the results of this ablation study, illustrating the relationship between the DSC and different values of γ. The scatter plot with a fitted curve (left) shows the relationship between DSC and various γ values, with optimal performance achieved at $\gamma = 0.5$. This suggests that an equal contribution from both prior knowledge and data-driven learning in the gating network yields the most accurate nuclei segmentation. To ensure the robustness and consistency of the model's performance across different initializations, we evaluated each configuration of γ over six independent runs, using random seeds ranging from 41 to 46. The box plot (right) in Fig. 7 displays the distribution of DSC scores for each γ value. The compact boxes and minimal outliers underscore the stability and reliability of MoE-NuSeg's performance, regardless of initialization.

Table 4
Comparison of model complexity and computational efficiency among MoE-NuSeg and the state-of-the-art methods. 'I' and 'II' refer to stage one and two of MoE-NuSeg, respectively. The table presents the number of parameters, FLOPs, and training/inference times per 100 images for each model. MoE-NuSeg strikes a balance between complexity and efficiency, with competitive FLOPs and inference time, demonstrating the effectiveness of the proposed architecture in optimizing performance while maintaining computational efficiency.

Models	#Params ($\times 10^6$)	FLOPs ($\times 10^9$)	Training/Inference (s)
SegNet	29.44	160.52	9.57/4.49
U-Net	10.55	42.74	13.09/2.65
TransUNet	66.82	130.41	11.09/3.16
Swin-Unet	27.15	31.13	12.84/0.74
CA$^{2.5}$-Net	24.27	460.29	8.27/0.68
TransNuSeg	29.99	193.93	23.53/1.56
MoE-NuSeg (I)	20.39	59.72	37.71/1.07
MoE-NuSeg (I+II)	20.95	116.01	13.61/2.36

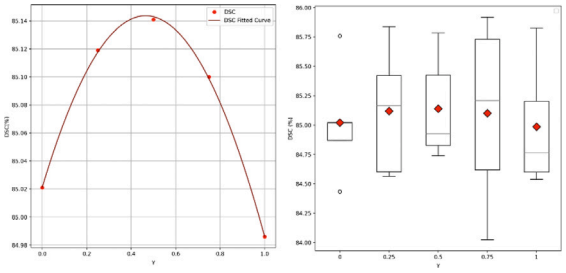

Fig. 7. Ablation study on the hyperparameter γ controls the balance between prior weights (w_{pre}) and data-driven weights (w_i) in the gating network. The left plot shows DSC across different γ values, with optimal performance achieved at $\gamma = 0.5$. The right box plot presents the distribution of DSC scores across six independent runs, demonstrating the model's consistency and robustness.

We then conducted an ablation study comparing the model's performance with single-stage (I) and two-stage (II) training. The results in Table 5 clearly demonstrate the effectiveness of the two-stage training approach over the single-stage setup. In the single-stage approach, the experts and gating network are trained concurrently, allowing them to learn and adapt together from the start. In contrast, the two-stage approach first trains the experts independently, enabling them to specialize in their respective tasks (nuclei segmentation, normal edge segmentation, and clustered edge segmentation). In the second stage, their collaboration with the gating network is fine-tuned. Across all datasets and evaluation metrics, the two-stage approach consistently yields better performance. For instance, on the Histology dataset, the model achieves a DSC of 85.13% ± 0.42% with two-stage training, compared to 84.76% ± 0.56% with single-stage training. Similar improvements are observed in F1 score, IoU, Accuracy, and Precision, where the two-stage training demonstrates a notable advantage.

The enhanced performance of the two-stage approach can be attributed to several factors. First, training the experts independently allows them to capture more refined and discriminative features without interference from the gating network, enhancing their performance in each subtask. Second, the fine-tuning stage enables effective collaboration between the experts and the gating network, leveraging the specialized knowledge gained from the first stage to make more informed decisions by combining expert outputs wisely. The consistent improvements across all datasets and metrics underscore the robustness and generalizability of the two-stage training scheme. By

Table 5
Ablation study comparing the performance of MoE-NuSeg with single-stage (I) and two-stage (II) training. Results are presented as mean ± 95% confidence intervals ($n = 6$), with best performance highlighted in **bold**.

Metrics	Stages	Histology	Fluorescence	Lizard
DSC (%)	I	84.76 ± 0.56	95.58 ± 0.37	74.86 ± 2.17
	II	**85.13** ± 0.42	**95.93** ± 0.39	**77.36** ± 1.12
F1 (%)	I	84.28 ± 3.02	96.28 ± 1.73	76.08 ± 1.29
	II	**84.71** ± 3.07	**96.67** ± 1.48	**77.09** ± 2.54
IoU (%)	I	73.66 ± 0.84	91.59 ± 0.61	60.85 ± 2.65
	II	**74.23** ± 0.63	**92.26** ± 0.67	**63.16** ± 1.48
Acc (%)	I	91.16 ± 0.39	96.42 ± 0.34	89.21 ± 0.84
	II	**91.38** ± 0.35	**96.73** ± 0.38	**90.44** ± 0.59
Precision (%)	I	84.35 ± 1.16	94.90 ± 0.81	73.38 ± 2.24
	II	**84.82** ± 1.00	**95.61** ± 0.65	**76.29** ± 0.90

Table 6
Ablation study demonstrating the statistical significance of the two-stage training scheme in MoE-NuSeg. The table presents the p-values from paired one-tailed t-tests comparing MoE-NuSeg's performance with and without the two-stage training. Consistently low p-values (<0.05) across all datasets and metrics provide strong evidence that the two-stage training improves the model's performance across all datasets.

Metrics	Histology	Fluorescence	Lizard
DSC (%)	0.011	0.005	0.011
F1 (%)	0.010	0.038	0.012
IoU (%)	0.012	0.003	0.010
Acc (%)	0.011	0.002	0.008
Precision (%)	0.075	0.001	0.009

allowing the experts to specialize independently and then fine-tuning their collaboration, MoE-NuSeg adapts effectively to various image modalities.

Finally, Table 6 shows the statistical significance of these improvements. The consistently low p-values across all datasets and metrics provide strong evidence of the performance gains achieved with the two-stage training scheme.

5. Limitations and future work

While MoE-NuSeg shows performance advantage in nuclei segmentation across diverse datasets, several limitations and opportunities for future research that should be acknowledged.

First, although MoE-NuSeg has been evaluated on five datasets, further validation on a broader range of biomedical images, particularly with different tissue types, would enhance its generalizability and robustness. Future studies should explore applying MoE-NuSeg to additional modalities, such as electron microscopy or multi-modal imaging, to assess its performance and adaptability. Second, while the attention-based gating network effectively integrates expert outputs, improving its interpretability would provide deeper insights into how the model adapts to different image characteristics. Developing more transparent gating strategies, such as attention visualization or interpretable attention mechanisms, could help elucidate the dynamic weighting of expert contributions. Third, the current focus of MoE-NuSeg is nuclei segmentation, with edge segmentation serving as auxiliary tasks. Expanding the architecture to include additional tasks, such as cell type classification or morphological feature extraction, could provide a more comprehensive analysis of cellular structures, enhancing its utility in specific research and clinical applications.

Another limitation is MoE-NuSeg's focus on 2D nuclei segmentation. Extending MoE-NuSeg to handle 3D volumetric data would be a valuable direction for future work, particularly for imaging techniques like confocal or electron microscopy. Adapting the model to 3D, such as through a 3D Swin Transformer, would allow the capture of spatial relationships across three dimensions. Given the resource-intensive nature of 3D data processing, incorporating memory-efficient mechanisms, such as flash-attention, could help accelerate computations.

Additionally, MoE-NuSeg currently employs a limited number of domain-specific experts. Future research could explore architectures with a larger pool of simpler experts, inspired by swarm intelligence. As the scale of experts increases exponentially, emergent behaviors or capabilities could appear, leading to more dynamic and adaptive solutions. This approach could allow experts to evolve over time, enhancing the model's adaptability without relying on complex configurations, while fostering efficient collaboration and self-organization among a large number of simple experts.

Lastly, integrating MoE-NuSeg with foundation models offers new possibilities. Leveraging pre-trained models on large, diverse datasets could enhance MoE-NuSeg's generalization to unseen tasks and imaging modalities. Using a foundation model as a backbone for expert training could enable MoE-NuSeg to dynamically adapt to a wide range of medical imaging tasks, benefiting from the vast knowledge embedded in foundation models.

6. Conclusion

In this study, we introduced MoE-NuSeg, a Mixture of Experts network for nuclei segmentation that integrates domain knowledge with an attention-based gating mechanism to achieve state-of-the-art performance. By leveraging multiple specialized experts and dynamically allocating their contributions, MoE-NuSeg demonstrates improved accuracy, robustness, and generalizability across diverse datasets and imaging modalities. MoE-NuSeg's key innovations lie in its domain-knowledge-driven experts and its efficient attention-based gating network, which balances specialized knowledge with parameter efficiency, allowing the model to adapt to complex nuclei segmentation tasks.

Experiments on three distinct datasets validate the effectiveness of MoE-NuSeg, consistently outperforming existing models. The two-stage training process, which first trains experts independently and then fine-tunes them with the gating network, further enhances the model's ability to integrate expert outputs and deliver optimal segmentation performance. By combining domain expertise and attention mechanisms within a Mixture of Experts framework, MoE-NuSeg provides insights for the field of biomedical image analysis. The methods and principles developed here can be adapted to other segmentation challenges, such as cell or organ segmentation, and offer potential for addressing broader tasks in medical image analysis.

CRediT authorship contribution statement

Xuening Wu: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Conceptualization. **Yiqing Shen:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Qing Zhao:** Visualization, Data curation. **Yanlan Kang:** Visualization, Data curation. **Wenqiang Zhang:** Supervision, Project administration, Methodology, Funding acquisition.

Informed consent statement

The study does not involve any human or animal study.

Fund statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available on request from the corresponding author.

References

[1] H. Irshad, A. Veillard, L. Roux, D. Racoceanu, Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential, IEEE Rev. Biomed. Eng. 7 (2013) 97–114.

[2] A. Lagree, M. Mohebpour, N. Meti, K. Saednia, F.-I. Lu, E. Slodkowska, S. Gandhi, E. Rakovitch, A. Shenfield, A. Sadeghi-Naini, et al., A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks, Sci. Rep. 11 (1) (2021) 8025.

[3] X. Li, C. Li, M.M. Rahaman, H. Sun, X. Li, J. Wu, Y. Yao, M. Grzegorzek, A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches, Artif. Intell. Rev. 55 (6) (2022) 4809–4878.

[4] E. Budginaitė, M. Morkūnas, A. Laurinavičius, P. Treigys, Deep learning model for cell nuclei segmentation and lymphocyte identification in whole slide histology images, Informatica (Ljubl.) 32 (1) (2021) 23–40.

[5] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[6] S. Lal, D. Das, K. Alabhya, A. Kanfade, A. Kumar, J. Kini, Nucleisegnet: Robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images, Comput. Biol. Med. 128 (2021) 104075.

[7] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[8] S. Wazir, M.M. Fraz, Histoseg: Quick attention with multi-loss function for multi-structure segmentation in digital histology images, in: 2022 12th International Conference on Pattern Recognition Systems, ICPRS, IEEE, 2022, pp. 1–7.

[9] P. Shrestha, N. Kuang, J. Yu, Efficient end-to-end learning for cell segmentation with machine generated weak annotations, Commun. Biol. 6 (1) (2023) 232.

[10] J. Deng, Y. Shen, Y. Guo, J. Ke, Cellsegnet: an adaptive multi-resolution hybrid network for cell segmentation, in: Medical Imaging 2022: Digital and Computational Pathology, Vol. 12039, SPIE, 2022, pp. 242–248.

[11] J. Ke, Y. Lu, Y. Shen, J. Zhu, Y. Zhou, J. Huang, J. Yao, X. Liang, Y. Guo, Z. Wei, et al., Clusterseg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets, Med. Image Anal. 85 (2023) 102758.

[12] H. Chen, X. Qi, L. Yu, P.-A. Heng, Dcan: deep contour-aware networks for accurate gland segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2487–2496.

[13] Y. Zhou, O.F. Onder, Q. Dou, E. Tsougenis, H. Chen, P.-A. Heng, Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation, in: Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26, Springer, 2019, pp. 682–693.

[14] J. Huang, Y. Shen, D. Shen, J. Ke, Ca 2.5-net nuclei segmentation framework with a microscopy cell benchmark collection, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, Springer, 2021, pp. 445–454.

[15] Z. He, M. Unberath, J. Ke, Y. Shen, Transnuseg: A lightweight multi-task transformer for nuclei segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 206–215.

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.

[17] C. Wang, R. Xu, S. Xu, W. Meng, X. Zhang, Da-net: Dual branch transformer and adaptive strip upsampling for retinal vessels segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 528–538.

[18] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Comput. 3 (1) (1991) 79–87.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[21] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[22] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[23] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[25] F. Hörst, Rempe, et al., Cellvit: Vision transformers for precise cell segmentation and classificationf, Med. Image Anal. 94 (2024) 103143.

[26] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[27] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[28] A. Kirillov, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[29] K. Zhang, D. Liu, Customized segment anything model for medical image segmentation, 2023, arXiv preprint arXiv:2304.13785.

[30] J. Cheng, et al., Sam-med2d, 2023, arXiv preprint arXiv:2308.16184.

[31] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, et al., Segment anything model for medical images? Med. Image Anal. 92 (2024) 103061.

[32] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nature Commun. 15 (1) (2024) 654.

[33] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, Y. Jin, Medical sam adapter: Adapting segment anything model for medical image segmentation, 2023, arXiv preprint arXiv:2304.12620.

[34] J.M.J. Valanarasu, V.M. Patel, Unext: Mlp-based rapid medical image segmentation network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 23–33.

[35] M.J. Saks, R.F. Kidd, Human information processing and adjudication: Trial by heuristics, Law Soc'y Rev. 15 (1980) 123.

[36] A.K. Shah, D.M. Oppenheimer, Heuristics made easy: an effort-reduction framework, Psychol. Bull. 134 (2) (2008) 207.

[37] S. Frintrop, E. Rome, H.I. Christensen, Computational visual attention systems and their cognitive foundations: A survey, ACM Trans. Appl. Percept. (TAP) 7 (1) (2010) 1–39.

[38] R. Verma, N. Kumar, A. Patil, N.C. Kurian, S. Rane, S. Graham, Q.D. Vu, M. Zwager, S.E.A. Raza, N. Rajpoot, et al., Monusac2020: A multi-organ nuclei segmentation and classification challenge, IEEE Trans. Med. Imaging 40 (12) (2021) 3413–3423.

[39] S. Graham, M. Jahanifar, A. Azam, M. Nimir, Y.-W. Tsang, K. Dodd, E. Hero, H. Sahota, A. Tank, K. Benes, et al., Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 684–693.

[40] F. Kromp, E. Bozsaky, F. Rifatbegovic, L. Fischer, M. Ambros, M. Berneder, T. Weiss, D. Lazic, W. Dörr, A. Hanbury, et al., An annotated fluorescence image dataset for training nuclear segmentation methods, Sci. Data 7 (1) (2020) 262.

[41] P. Naylor, M. La, F. Reyal, T. Walter, Segmentation of nuclei in histopathology images by deep regression of the distance map, IEEE Trans. Med. Imaging 38 (2) (2019) 448–459.

[42] Q.D. Vu, S. Graham, T. Kurc, M.N.N. To, M. Shaban, T. Qaiser, N.A. Koohbanani, S.A. Khurram, J. Kalpathy-Cramer, T. Zhao, et al., Methods for segmentation and classification of digital microscopy tissue images, Front. Bioeng. Biotechnol. 7 (2019) 433738.

[43] Y. Shen, Y. Luo, D. Shen, J. Ke, Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 212–221.

[44] Y. Shen, J. Ke, Staindiff: Transfer stain styles of histology images with denoising diffusion probabilistic models and self-ensemble, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 549–559.

[45] D. Müller, I. Soto-Rey, F. Kramer, Towards a guideline for evaluation metrics in medical image segmentation, BMC Res. Notes 15 (1) (2022) 210.

[46] H. Hsu, P.A. Lachenbruch, Paired t test, in: Wiley StatsRef: Statistics Reference Online, 2014.

[47] R.R. Bahadur, A property of the t-statistic, Sankhyā (1952) 79–88.

[48] H. Kim, H. Yoon, N. Thakur, G. Hwang, E.J. Lee, C. Kim, Y. Chong, Deep learning-based histopathological segmentation for whole slide images of colorectal cancer in a compressed domain, Sci. Rep. 11 (1) (2021) 22520.