Journal of Multimedia Information System VOL. 11, NO. 4, December 2024 (pp. 219-228): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2024.11.4.219

Enhancing Nuclei Segmentation in Histology Images by Effectively Merging Task-Specific and Foundation Models

Bishal Ranjan Swain[1], Kyung Joo Cheoi[2], Jaepil Ko[1*]

Abstract

Nuclei segmentation in histology images is a critical task for various diagnostic applications in digital pathology. Traditional task-specific models, such as U-Net based architectures, are adept at capturing fine-grained and local information but often struggle with integrating global contextual features. On the other hand, foundation models effectively capture coarse-level global context but lack the precision required for detailed segmentation tasks. In this work, we propose a novel approach to bridge this gap by merging the strengths of task-specific and foundation models. We introduce a Gated Fusion Block that leverages the task-specific capabilities of U-Net-like architectures and merges with generalizable knowledge from foundation models. It combines local and global representations through gated squeeze-and-excitation layer followed by adaptive feature selection and cross-attention. We demonstrate the effectiveness of this approach through extensive experimentation on multiple histology datasets. The results show significant improvements in segmentation performance, with a 12% and 17.22% increase in Dice score and mIoU, respectively, on the CryoNuSeg dataset, a 15.55% and 16.77% improvement on the NuInsSeg dataset, and a 9% improvement in both metrics for the CoNIC dataset. Our findings highlight the potential of integrating task-specific and foundation models to achieve state-of-the-art results in nuclei segmentation.

Key Words: Histopathology Segmentation, Nuclei Segmentation, Image Segmentation, Medical Imaging.

I. INTRODUCTION

In digital pathology, accurate segmentation of nuclei in histological images is vital for cancer diagnosis and prognosis [1]. Segmentation of nuclei enables detailed examination of cellular structures and behaviors, such as analyzing cell cycles, mutations, and the morphology of cancerous tissues [2]. These analyses are essential for identifying cancer types, evaluating severity, and guiding treatment plans [3]. Tissue biopsy-based diagnostics relies heavily on this type of segmentation and remains the gold standard for cancer detection worldwide [4]. Given the volume and complexity of biopsy samples, automated and accurate segmentation methods have become critical to aid pathologists and streamline diagnostic workflows.

Nuclei segmentation is a fundamental step in quantitative histopathology that enables the analysis of cellular morphology, spatial distribution, and tissue architecture [4]. Accurate segmentation is crucial for diagnosing various diseases, including cancer, where nuclear features are key indicators of malignancy [5]. Traditional manual annotation is time-consuming and subject to inter-observer variability [6], highlighting the need for automated solutions.

Task-specific models, such as the U-Net family of architectures, have emerged as the benchmark for medical image segmentation [5]. U-Net [6] variants are effective in capturing the fine-grained details that are necessary for high-precision cell segmentation. They are effective due to their encoder-decoder structure and skip connections which help preserve contextual information. Despite their success, task-specific models typically require large amounts of labeled training data and often struggle to generalize across different staining variations and imaging modalities that are commonly encountered in histopathological datasets [7]. Moreover, these models tend to focus on localized features and may not fully capture the broader context of tissue structures, which is crucial for understanding complex cell and tissue interactions [8-9]. Moreover, challenges such as heterogeneity in staining, presence of artifacts, and variability in nuclear morphology necessitate more robust approaches.

To address the above limitations, recent research has ex-

Manuscript received October 23, 2024; Revised December 03, 2024; Accepted December 09, 2024. (ID No. JMIS-24M-10-031)
Corresponding Author (*): Jaepil Ko, +82-54-478-7529, nonezero@kumoh.ac.kr
[1]Department of Computer Engineering, Kumoh National Institute of Technology, Gumi, Korea, bishalswain@kumoh.ac.kr, nonezero@kumoh.ac.kr
[2]Department of Computer Science, Chungbuk National University, Cheongju, Korea, kjcheoi@chungbuk.ac.kr

plored the potential of foundation models, such as the Segment Anything Model (SAM) [10], which are trained on vast and diverse datasets. Foundation models excel in capturing global contextual features and generalize well across various visual domains. However, when applied to histology images, foundational models often lack the pixel-level precision needed for accurate pathological analysis, especially in segmenting small, densely packed nuclei [7]. Our study aims to address these challenges by integrating global contextual information from foundation models with the precision of task-specific models.

In this paper, we propose a novel approach that integrates the strengths of task-specific and foundation models to enhance nuclei segmentation in histology images. We enhance the U-Net3+ architecture [11] by introducing an adaptive feature selection mechanism, which we call, *e*U-Net3+. Additionally, we propose an Enhanced Fusion Block (EFB), which dynamically fuses the global contextual knowledge from foundation models with the detailed local representations from task-specific models using cross-attention and gated squeeze-and-excitation techniques [12]. Our proposed framework enables the model to leverage both global context and local precision, addressing the challenges posed by complex histological images.

Our approach demonstrates significant improvements in segmentation performance, achieving a 12% and 17.22% increase in Dice score and mIoU, respectively, on the CryoNuSeg dataset [13], a 15.55% and 16.77% increase on the NuInsSeg dataset [14], and a 9% improvement on both metrics for the CoNIC dataset [15]. By effectively merging task-specific models with foundation models, we set a new standard for state-of-the-art nuclei segmentation in digital pathology. The main contributions of this paper are as follows:

- **Integration of Task-Specific and Foundation Models**: We propose a framework that effectively combines the fine-grained feature extraction capabilities of task-specific models with the global contextual understanding of foundation models. Our Enhanced Fusion Block (EFB) dynamically fuses local and global features through cross-attention and gated squeeze-and-excitation techniques.
- **Adaptive Feature Selection using GLUs:** We introduce an adaptive feature selection mechanism using Gated Linear Units (GLUs) within the U-Net3+ architecture to create *e*U-Net3+, which enhances local feature extraction and improves segmentation accuracy.
- **State-of-the-Art Nuclei Segmentation Results**: By setting new benchmarks on the CryoNuSeg, NuInsSeg, and CoNIC datasets, our work advances the field of nuclei segmentation in digital pathology.

II. RELATED WORKS

Nuclei segmentation in histopathological images has witnessed significant advances with the advent of deep learning, particularly convolutional neural networks (CNNs) [6]. Among CNN-based architectures, U-Net has become a seminal model for medical image segmentation due to its symmetrical encoder-decoder structure and the use of skip connections, which preserve spatial details while extracting higher-level features [16]. Variants of U-Net [16-19] have demonstrated remarkable success in various biomedical tasks, including nuclei segmentation, by improving multi-scale feature extraction and incorporating attention mechanisms.

Recent approaches have focused on hybrid models that combine U-Net with other architectures to leverage the strengths of both [20]. These hybrid approaches address the limitations of U-Net in handling complex visual features like overlapping nuclei, irregular shapes, and varying sizes. For instance, ASPPU-Net [21] integrates Atrous Spatial Pyramid Pooling (ASPP) with U-Net, enhancing its ability to capture multi-scale contextual information. Similarly, Hover-Net [22] extends U-Net with residual connections and dense blocks to improve feature reuse and boundary precision, particularly in dense cell regions. Another notable advancement is the Sharp U-Net models [23], which aim to increase performance by minimizing low-frequency noise introduced during down-sampling and up-sampling layers. Attention mechanisms have also become a popular enhancement to U-Net variants. DEAU [24] introduces an Attention Encoding Path (AEP) that runs parallel to the U-Net's traditional encoding path, refining feature extraction by using attention maps that prioritize diagnostically significant regions. Similarly, methods incorporating self-attention or cross-attention layers have shown improved nuclei detection and segmentation accuracy by capturing long-range dependencies and suppressing irrelevant background information. DDU-Net [25] leverages dual decoders to handle both nuclear and cytoplasmic regions, enhancing segmentation performance on histopathological images where overlapping structures are prevalent. U-Net3+ [11], on the other hand, focuses on incorporating full-scale skip connections and deep supervision to better fuse multi-scale features. To benchmark our proposed model, DDU-Net and U-Net3+ were selected as baseline models as they previously obtained state-of-the-art results. Both architectures are well-regarded for their effectiveness in medical image segmentation. While U-Net3+ excels in fusing full-scale feature maps and providing deep supervision, DDU-Net's dual-decoder design allows it to effectively differentiate between different cellular structures, making these models strong candidates for comparison.

Additionally, some studies have employed wavelet-based channel attention modules to capture more global context within U-Net-based models [26]. By decomposing feature maps into different frequency components using wavelet transforms can help these methods effectively focus on salient features at various scales and enhance segmentation performance by integrating both local and global information. Post-processing techniques have also been explored to improve segmentation results. Studies like [27] have used morphological operations to refine segmentation masks. Semi-supervised [28-30] and unsupervised learning approaches [31-34] have also been investigated for nuclei segmentation, aiming to reduce the reliance on large amounts of annotated data. Methods such as [35,36] utilize generative adversarial networks (GANs) or self-supervised learning techniques to learn representations from unlabeled data. However, these methods often struggle to achieve high performance due to the complexity and variability of histopathological images, and the lack of explicit guidance from labeled examples limits their effectiveness compared to fully supervised approaches.

Another trend is the integration of foundation models such as Vision Transformers (ViTs) [37] that are pre-trained on vast and diverse datasets. Foundation models offer strong generalization across various domains by learning rich, global contextual features, as demonstrated by the Segment Anything Model (SAM) [10]. While these models excel at capturing global context, their application in medical image segmentation, particularly nuclei segmentation, requires finetuning or parameter-based optimizations. SAM is trained on natural images that lacks the fine-grained detail necessary for precise nuclei boundary detection, making it necessary to finetune or adapt it to task-specific demand like nuclei segmentation. There have been recent developments that include MedSAM [38] which is specifically trained on medical images. But even then, it is far from reaching the performance level of task-specific models semantic segmentation models.

Our work builds on these developments by proposing a novel hybrid model that combines the global context awareness of SAM with the fine-grained feature extraction capabilities of U-Net. By incorporating GLU and GFB, our approach dynamically and effectively fuses local and global features, allowing the model to address the challenges of nuclei segmentation in complex histopathological images. While many previous models focus on either task-specific or general foundation models, our approach effectively combines both to achieve state-of-the-art performance, as demonstrated by our significant improvements in Dice scores across multiple datasets.

III. METHODS

The overall pipeline of the proposed approach is depicted in Fig. 1. Our methodology focuses on enhancing task-specific nuclei segmentation by leveraging the combined strengths of a task-specific model and a foundation model. The task-specific model is an enhanced version of U-Net3+ (referred to as *e*U-Net3+), optimized for fine-grained feature extraction. In parallel, a pre-trained Segment Anything Model (SAM) provides global contextual information. These local and global features are fused through our proposed Enhanced Fusion Block (EFB), which combines Gated Linear Units (GLUs) in a squeeze-and-excitation mechanism followed by a cross-attention block. This ensures effective integration of both global and local representations for enhanced segmentation performance.

3.1. Task Specific Model

The U-Net architecture has emerged as a fundamental framework in medical image segmentation due to its unique ability to balance both high-level semantic information and low-level spatial details [20]. Over time, numerous variants of the U-Net architecture have been proposed to address specific challenges in medical image segmentation. Most variations of U-Net introduce enhancements to tackle particular issues such as multi-scale feature extraction and deeper supervision. From among the U-Net based models, U-Net3+ introduces full scale skip connections and allows

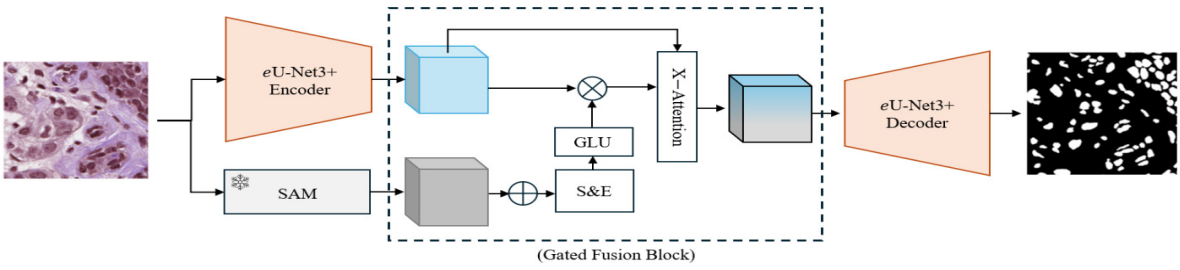


Fig. 1. Architecture of the proposed model. The S&E is squeeze and excitation block, GLU is gated linear unit block and X-attention is cross attention block.

features from all levels of the encoder to be directly connected to the corresponding layers in the decoder. This deep supervision improves gradient flow during training and facilitates better feature learning by incorporating multiple output layers. While U-Net3+ provides a strong foundation for segmenting histopathological images, traditional activation functions such as ReLU [39] apply uniform transformations to all features and overlook the nuances in densely packed or morphologically diverse regions. To address this limitation, we enhance U-Net3+ by GLUs, allowing for adaptive feature selection.

GLUs split the input into two streams: one undergoes a linear transformation, while the other passes through a sigmoid activation function. The sigmoid activation serves as a gate, modulating the flow of information based on the relevance of each feature to the segmentation task. This mechanism allows the network to selectively focus on diagnostically significant regions, such as densely packed nuclei, while ignoring less relevant features. The GLU operation is defined as:

$$GLU(x) = sigmoid(W_a \cdot x + b_a) \odot (W_l \cdot b_l), \qquad (1)$$

where, $\odot$ denotes element-wise multiplication, W_a, b_a represent the weights and biases for gating mechanism, and W_l, b_l represent the weights and biases for the linear transformation. This enhancement ensures that the task-specific model not only captures multi-scale features but also fine-tunes its focus towards the most significant regions in the image.

3.2. Foundational Model

The Segment Anything Model (SAM) [10] is used as a foundation model. SAM is pre-trained on a vast dataset comprising over 11 million images and one billion masks. Although SAM is not specialized for medical images, its robust ability to capture high-order global context across diverse visual domains makes it a valuable asset for guiding nuclei segmentation. SAM's ability to identify global contextual relationships aids the task-specific model in understanding broader tissue structures, which is particularly helpful when analyzing complex histological images.

In our proposed framework, we use SAM's Base model checkpoint (91M parameters) – ViT B, as it provides a more abstract and ambiguous representation of the input image, which complements the detailed focus of eU-Net3+ as can be seen in Fig. 2. This ambiguity allows the task-specific model to fine-tune its decisions, particularly in differentiating nuclei from surrounding regions. Visualizing SAM's encodings through Principal Component Analysis (PCA), we observe that simpler SAM representations produced by the base ViT enable a better fusion with the local features captured by eU-Net3+.

3.3. X-Gated Fusion Block

A key challenge when combining feature representations from task-specific and foundation models is the potential for conflicting or redundant information. Simple concatenation of features often results in suboptimal performance due to this misalignment [40]. To address this, we introduce Enhanced Fusion Block (EFB) to effectively integrate the global context with local task-specific features.

The EFB consists of three main components: a Gated Squeeze-and-Excitation block, GLU block and a Cross-Attention Block. The squeeze-and-excitation block starts by performing adaptive average pooling on the concatenated features from SAM and eU-Net3+. This compresses spatial dimensions to focus on global information from each channel. The gated mechanism, implemented using GLUs, selectively allows important features to flow through, emphasizing only the most relevant global and local features. The cross-attention block operates on the gated features by treating them as queries, keys, and values in a standard attention operation. This mechanism enhances the model's ability to highlight important features, suppress irrelevant information, and increase contextual awareness.

IV. EXPERIMENTS AND RESULTS

4.1. Datasets

We used three publicly available histopathological d atasets that contained variability in staining techniques, and tissue morphology.

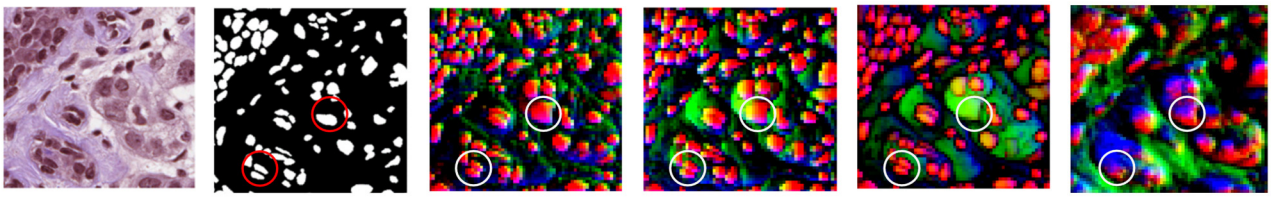


Fig. 2. Displays the representations of the encoded image of various SAM models. ViT B denotes SAM with ViT Base image weights, ViT L for Large image weights and H for Huge image weights. MedSAM is the adapted variant of SAM trained on medical images. The red regions denote the nuclei regions and green regions denote membrane. It is seen that ViT B seem to have more potential nuclei regions compared to other frozen models.

- **CryoNuSeg [13]:** This dataset consists of 30 high-resolution images obtained using cryo-sectioning techniques. It is known for its variability in nuclear morphology and presents challenges related to segmentation accuracy due to its complex staining and tissue structures.
- **NuInsSeg [14]:** Comprising 665 image patches, NuInsSeg is a challenging dataset characterized by diverse tissue types and staining methods. The dataset includes densely packed nuclei and complex tissue structures, making it an ideal test bed for assessing segmentation performance on intricate regions.
- **CoNIC [15]:** The CoNIC dataset includes 4,981 images and presents one of the largest and most complex datasets used for nuclei segmentation. It contains multiple tissue types and wide variations in nuclear shapes, providing a robust benchmark for evaluating generalization across diverse histological samples.

4.2. Experiment Setup

Our experiments were conducted using the PyTorch framework on a system equipped with an NVIDIA RTX A6000 GPU. To ensure consistency and standardization, all images were resized to a resolution of 256×256 pixels. The training process spanned 50 epochs, with an initial learning rate of 1×10^{-4} and a batch size of 16. We employed the Adam optimizer [41], which is well-suited for segmentation tasks, and incorporated a dropout rate of 0.3 to prevent overfitting.

All the images were preprocessed by using stain normalization as described in [42] and then augmentations were performed. We applied a combination of photometric and geometric augmentations as described in [43]. Geometric augmentations such as rotations, scaling, flips were performed to obtain samples from different perspectives and scales. Additionally, elastic deformations that mimic the natural deformations in biological tissues were performed that helped in enhancing the model's ability to handle non-rigid transformations and complex variations. Photometric augmentations like Gamma and intensity level transformations along with contrast limited adaptive histogram equalization were performed. These augmentations aimed to increase the diversity of the training set and help the model generalize better to unseen data.

For the loss function, we used a weighted combination of Dice loss [43] and focal loss [45], each contributing equally to the overall loss function. Dice loss was chosen for its effectiveness in handling imbalanced datasets, particularly in scenarios where nuclei occupy a small portion of the image. Focal loss further helps by focusing on hard-to-segment regions, ensuring that rare and challenging nuclei instances are not overlooked during training.

To evaluate the segmentation performance, we used the Dice coefficient and mean Intersection over Union (mIoU), both of which are standard metrics in medical image segmentation tasks. These metrics provide a robust evaluation of the overlap between the predicted segmentation masks and the ground truth, with higher values indicating better segmentation quality.

4.3. Results

The experimental results are summarized in Table 1. It demonstrates that integrating SAM's global contextual features significantly improves the performance of the task-specific eU-Net3+ model across all datasets.

- **CryoNuSeg:** Our model achieved a 12% increase in Dice score and a 17.22% increase in mIoU compared to baseline models. The inclusion of SAM helped mitigate the effects of freezing artifacts by providing additional context to differentiate nuclei from artifacts.
- **NuInsSeg:** We observed over 15% improvement in Dice score and mIoU. The model effectively handled densely packed and overlapping nuclei by leveraging global contextual information from SAM, aiding in

Table 1. Model performance across different datasets.

Dataset	Model	Dice	mIoU
CryoNuSeg	U-Net	0.7371	0.610
	DDU-Net	0.8143	0.6822
	U-Net3+	0.778	0.6432
	AWGUNET	0.764	0.6381
	*e*U-Net3+	0.8401	0.7644
	SAM +*e*U-Net3+	**0.8942**	**0.8164**
NuInsSeg	U-Net	0.7997	0.6781
	DDU-Net	0.7154	0.6133
	U-Net3+	0.7844	0.7261
	AWGUNET	0.7792	0.7236
	*e*U-Net3+	0.8307	0.8163
	SAM +*e*U-Net3+	**0.9399**	**0.8938**
CoNIC	U-Net	0.7353	0.6214
	DDU-Net	0.827	0.7347
	U-Net3+	0.8474	0.7992
	AWGUNET	0.8419	0.7944
	*e*U-Net3+	0.8966	0.8539
	SAM +*e*U-Net3+	**0.9351**	**0.8869**

distinguishing individual nuclei in crowded regions.

- **CoNIC:** The model showed a 9% improvement in both Dice score and mIoU, demonstrating its ability to generalize across diverse tissue types despite significant variability in nuclear appearance.

These performance differences across datasets can be attributed to the unique characteristics and challenges presented by each dataset. Our proposed methodology even outperforms the recently released AWGUNET [26] that uses wavelet for guiding task-specific U-Net model.

The choice of activation function has a substantial impact on segmentation performance and is detailed in Table 2. The inclusion of GLUs as activation function in the *e*U-Net3+ architecture has significantly outperformed other commonly used activation functions, such as ReLU, LeakyReLU [46], and Swish [47], across all datasets.

The GLU-enhanced model achieved a Dice score of 0.8401 on CryoNuSeg, 0.8307 on NuInsSeg, and 0.8966 on CoNIC, surpassing ReLU, LeakyReLU, and Swish by notable margins. This demonstrates the superiority of GLU in enabling selective feature activation, allowing the model to focus on the most relevant regions of the image for precise segmentation, especially in complex histological samples. The gating mechanism of GLU provides an adaptive feature selection that dynamically activates diagnostically important features, leading to more refined segmentation results.

The effectiveness of the proposed Gated Fusion Block (GFB) is evident from the performance gains observed. As shown in Table 3, incorporating the GFB into the model resulted in marked improvements across all datasets. The eU-Net3+ model with GFB achieved higher Dice scores and mIoU values compared to the version without GFB and shows the importance of effective feature fusion in enhancing segmentation accuracy. The GFB effectively addressed the challenge of fusing local and global features by employing GLUs for selective gating and a cross-attention mechanism for feature alignment. This allowed the model to dynamically balance fine-grained local features and broader contextual insights, leading to more accurate segmentations.

In addition, we evaluated different variants of the SAM model. As shown in Table 4, the ViT-B (Base) model consistently outperformed the larger ViT-L and ViT-H models. The ViT-B model provided an optimal level of global context without overwhelming the task-specific model with excessive detail, allowing for better integration and improved segmentation results.

Table 2. Effectiveness of GLU as activation function.

Activation functions	CryoNuSeg	NuInsSeg	CoNIC
ReLU	0.778	0.7844	0.8474
LeakyReLU	0.7931	0.8012	0.8567
Swish	0.7583	0.7721	0.8439
GLU	**0.8401**	**0.8307**	**0.8966**

Table 3. Effectiveness of the proposed gated fusion block (GFB).

Model	CryoNuSeg	NuInsSeg	CoNIC
*e*U-Net3+	0.8401	0.8307	0.8966
*e*U-Net3+ w/o GFB	0.8235	0.8146	0.8918
***e*U-Net3+w/GFB**	**0.8942**	**0.9399**	**0.9351**

The primary reason for the superior performance of the ViT-B (Base) model can be attributed to its ability to maintain a more balanced level of ambiguity in its representations. Although the ViT-H (Huge) model, with its 636M parameters, provides more detailed and nuanced representations of the input images, can lead to the loss of necessary ambiguity in certain regions of the histological images, as illustrated in Fig. 2. The ViT-B model, with 91M parameters, provided a more optimal level of ambiguity in global context which allowed the *e*U-Net3+ task-specific model to make finer decisions in ambiguous regions, such as distinguishing between nuclei and non-nuclei areas. By allowing the task-specific model to handle these more nuanced decisions, rather than relying entirely on SAM's global representation, the proposed model was able to achieve more accurate segmentation results.

Fig. 3 visually depicts the performance of various models across datasets. The violin plots in Fig. 4 provide further insight into the distribution of Dice scores for models with and without SAM integration. Mean Dice Scores (dotted lines) show a clear improvement with SAM integration across all datasets.

The violin plot in Fig. 4 also illustrates the variance in Dice scores, showing that SAM-guided models produce more consistent and reliable results, with less variance in their predictions.

Fig. 5 shows the qualitative assessment of the proposed model. The circles areas in the image highlights some of the regions where inclusion of SAM in *e*U-Net3+ performed better than using only the task-specific model. The SAM

Table 4. Effectiveness of the SAM frozen models.

Frozen models	CryoNuSeg	NuInsSeg	CoNIC
ViT - B	**0.8942**	**0.9399**	**0.9351**
ViT - L	0.8745	0.8918	0.9072
ViT - H	0.8731	0.8867	0.8917

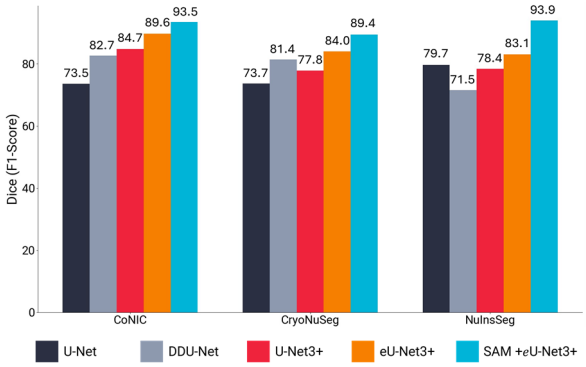


Fig. 3. Visual comparison of proposed model performance.

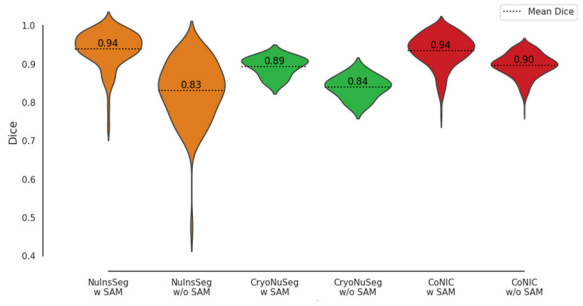


Fig. 4. Violin plot of dice score distribution during inference for with and without SAM integration.

guided predictions generally show more precise and accurate segmentation, further confirming the effects of the proposed method of adding global context from SAM to improve the segmentation accuracy of the task-specific model.

There are performance differences between three different datasets as the images in the datasets contain varied properties. In the data collection of histopathology images various techniques are used for creating the whole slide images. For instance, **CryoNuSeg** contains cryo-sectioned images that may include freezing artifacts affecting image clarity; **NuInsSeg** features densely packed nuclei from diverse tissues with varying staining techniques, making segmentation challenging; and **CoNIC** encompasses images from multiple organs with different staining protocols, resulting in significant variability in nuclear appearance. These differences contribute to the varied performance of our model across the datasets.

V. CONCLUSION

In this paper, we introduced a novel method for nuclei segmentation in histopathological images. We integrated task-specific models with foundation models to enhance performance. Specifically, we improved the U-Net3+ archi-

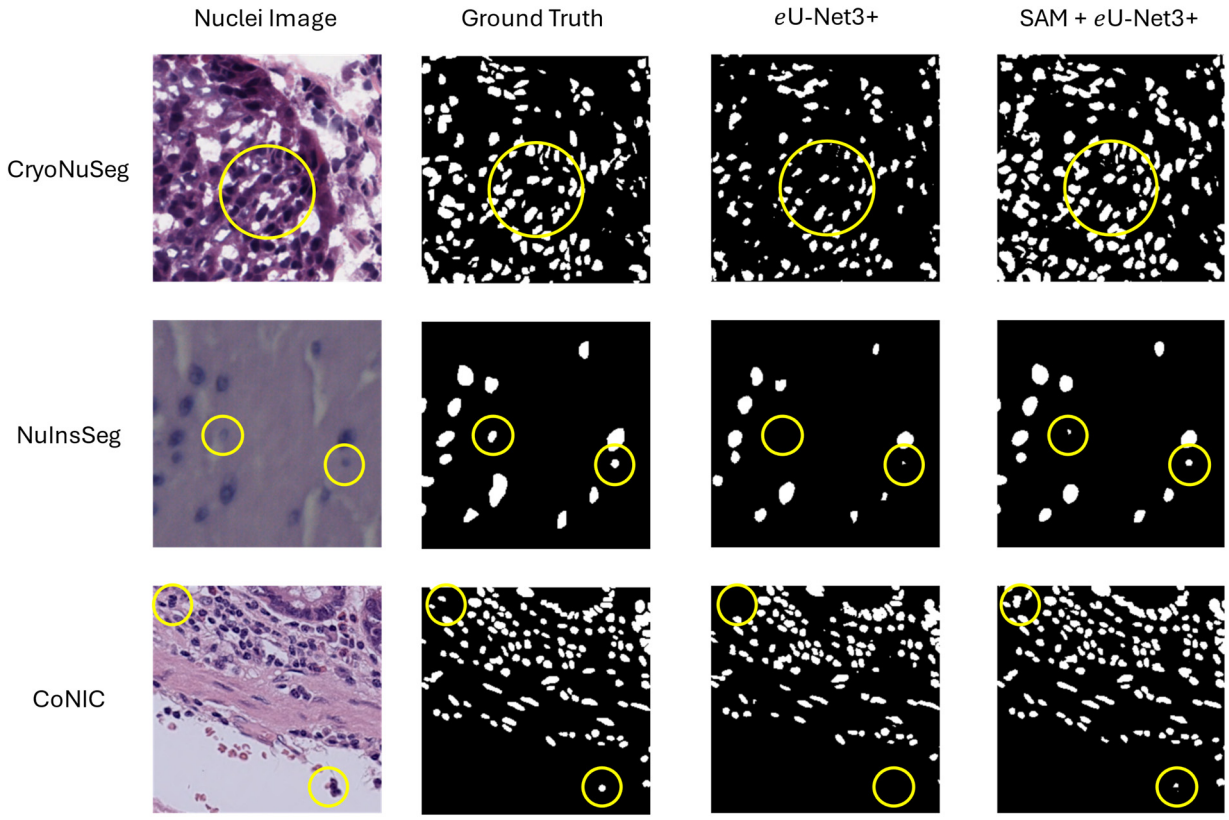


Fig. 5. Qualitative evaluation of segmentation performance of sample images across three datasets. ***e***U-Net3+ denotes the model with just adaptive feature selection and SAM+***e***U-Net3+ denotes the model with the SAM guidance. SAM+***e***U-Net3+ model shows more resemblance with the ground truth segmentation mask as highlighted in the images.

tecture by incorporating GLUs for adaptive feature selection. We also proposed GFB that dynamically fuses local (from *e*U-Net3+) and global representations (from SAM) using cross-attention and gated squeeze-and-excitation techniques. This fusion of global and local features enabled our model to tackle challenges such as varying tissue structures, complex staining techniques, and densely packed nuclei. Our experiments on three challenging histopathological datasets—CryoNuSeg, NuInsSeg, and CoNIC—demon-strated significant improvements in segmentation accuracy. Incorporating SAM's global context led to Dice score improvements of up to 12% on CryoNuSeg, 15.55% on NuInsSeg, and 9% on CoNIC compared to baseline models. Our analysis revealed that the ViT-B variant of SAM outperformed larger ViT models. It provided an optimal balance between capturing global context and computational efficiency. Visual assessments confirmed that SAM-guided models produced more accurate and reliable segmentations with reduced variance in performance. By effectively merging task-specific models with foundation models, our approach obtains the state-of-the-art results in nuclei segmentation. It also presents a flexible framework for integrating local and global representations in medical imaging. Future research can explore further optimizations of the fusion mechanisms. This includes experimenting with different attention strategies or adaptive weighting between local and global features.

ACKNOWLEDGMENT

This research was supported by Kumoh National Institute of Technology (2022−2024).

REFERENCES

[1] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550-1560, Jul. 2017.

[2] S. M. Gross, F. Mohammadi, C. Sanchez-Aguila, P. J. Zhan, T. A. Liby, and M. A. Dane, et al., "Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects," *Nature Communications*, vol. 14, no. 1, p. 3450, Jun. 2023.

[3] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers (Basel)*, vol. 15, no. 14, Jul. 2023.

[4] S. Wang, R. Rong, Q. Zhou, D. M. Yang, X. Zhang, and X. Zhan, et al., "Deep learning of cell spatial organizations identifies clinically relevant insights in tissue images," *Nature Communications*, vol. 14, no. 1, p. 7872, Dec. 2023.

[5] H. Li, J. Zhong, L. Lin, Y. Chen, and P. Shi, "Semi-supervised nuclei segmentation based on multi-edge features fusion attention network," *PLOS ONE*, vol. 18, no. 5, p. e0286161, May 2023.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[7] J. Ryu, A. V. Puche, J. Shin, S. Park, B. Brattoli, and J. Lee, et al., "Ocelot: Overlapped cell on tissue dataset for histopathology," 2023.

[8] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," 2022.

[9] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, and I. Rekik, et al., "Foundational models in medical imaging: A comprehensive survey and future vision," 2023.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, and L. Gustafson, et al., "Segment anything," 2023.

[11] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, and Y. Iwamoto, et al., "Unet 3+: A full-scale connected unet for medical image segmentation," 2020.

[12] A. Bhuiyan, Y. Liu, P. Siva, M. Javan, I. B. Ayed, and E. Granger, "Pose guided gated fusion for person re-identification," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, Mar. 2020, pp. 2664-2673.

[13] A. Mahbod, G. Schaefer, B. Bancher, C. Löw, G. Dorffner, and R. Ecker, et al., "Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned H&E stained histological images," *Computers in Biology and Medicine*, vol. 132, p. 104349, May 2021.

[14] A. Mahbod, C. Polak, K. Feldmann, R. Khan, K. Gelles, and G. Dorffner, et al., "Nuinsseg: A fully annotated dataset for nuclei instance segmentation in h&e-stained histological images," *Scientific Data,* vol. 11, no. 2, p. 295, 2023.

[15] S. Graham, M. Jahanifar, Q. D. Vu, G. Hadjigeorghiou, T. Leech, and D. Snead, et al., "Conic: Colon nuclei identification and counting challenge 2022," *arXiv Prep. arXiv:2111.14485,* 2021.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, and Y. Wang, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv Prep. arXiv:2102.04306,* 2021.

[17] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, and K. Misawa, et al., "Attention u-net: Learning where to look for the pancreas," *Prep. arXiv:1804.03999,* 2018.

[18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern*

Recognition, vol. 106, p. 107404, 2020.

[19] I. Kiran, B. Raza, A. Ijaz, and M. A. Khan, "Denseres-unet: Segmentation of overlapped/clustered nuclei from multi-organ histopathology images," *Computers in Biology and Medicine*, vol. 143, p. 105267, 2022.

[20] M. Traoré, E. Hancer, R. Samet, Z. Yıldırım, and N. Nemati, "Compsegnet: An enhanced u-shaped architecture for nuclei segmentation in H&E histopathology images," *Biomedical Signal Processing and Control*, vol. 97, p. 106699, 2024.

[21] T. Wan, L. Zhao, H. Feng, D. Li, C. Tong, and Z. Qin, "Robust nuclei segmentation in histopathology using asppu-net and boundary refinement," *Neurocomputing*, vol. 408, pp. 144-156, 2020.

[22] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, and J. T. Kwak, et al., "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, vol. 58, p. 101563, 2019.

[23] P. Senapati, A. Basu, M. Deb, and K. G. Dhal, "Sharp dense u-net: An enhanced dense u-net architecture for nucleus segmentation," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 6, pp. 2079-2094, Jun. 2024.

[24] A. Vahadane, B. Atheeth, and M. Shantanu, "Dual encoder attention U-Net for nuclei segmentation," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.

[25] Y. Wang, Y. Peng, W. Li, G. C. Alexandropoulos, J. Yu, and D. Ge, et al., "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-12, 2022.

[26] A. Roy, P. Pramanik, D. Kaplun, S. Antonov, and R. Sarkar, "AWGUNET: Attention-aided wavelet guided u-net for nuclei segmentation in histopathology images," *arXiv Prep. arXiv:2406.08425*, 2024.

[27] M. Gour, S. Jain, and T. S. Kumar, "Robust nuclei segmentation with encoder-decoder network from the histopathological images," *International Journal of Imaging Systems and Technology*, vol. 34, no. 4, p. e23111, 2024.

[28] C. Cui, R. Deng, Q. Liu, T. Yao, S. Bao, and L. W. Remedios, et al., "All-in-sam: From weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning," *Journal of Physics: Conference Series*, vol. 2722, no. 1, p. 012012.

[29] Y. Zhang, Z. Wang, Y. Wang, H. Bian, L. Cai, and H. Li, et al., "Boundary-aware contrastive learning for semi-supervised nuclei instance segmentation," *arXiv Prep. arXiv:2402.04756*, 2024.

[30] J. Ren, J. Che, P. Gong, X. Wang, X. Li, and A. Li, et al., "Cross comparison representation learning for semi-supervised segmentation of cellular nuclei in immunofluorescence staining," *Computers in Biology and Medicine*, vol. 171, p. 108102, 2024.

[31] Q. Zhang, Z. Ying, J. Shen, S. K. Kou, J. Sun, and B. Zhang, "Unsupervised color-based nuclei segmentation in histopathology images with various color spaces and K values selection," *International Journal of Image and Graphics*, p. 2550061, 2024.

[32] Z. Ahmed, C. N. E. A. Siddiqi, F. F. Alam, T. Ahmed, and T. M. Chowdhury, "Nuclei instance segmentation of cryosectioned H&E stained histological images using triple U-Net architecture," *arXiv Prep. arXiv: 2404.12986,* 2024.

[33] X. Yang, B. Ding, J. Qin, L. Guo, J. Zhao, and Y. He, "HVS-Unsup: Unsupervised cervical cell instance segmentation method based on human visual simulation," *Computers in Biology and Medicine*, vol. 171, p. 108147, 2024.

[34] B. Kochetov, P. D. Bell, P. S. Garcia, A. S. Shalaby, R. Raphael, and B. Raymond, et al., "UNSEG: Unsupervised segmentation of cells and their nuclei in complex tissue samples," *Communications Biology*, vol. 7, no. 1, p. 1062, 2024.

[35] S. Xu, G. Li, H. Song, J. Wang, Y. Wang, and Q. Li, "GeNSeg-Net: A general segmentation framework for any nucleus in immunohistochemistry images," in *Proceedings of the 32nd ACM International Conference on Multimedia,* 2024, pp. 4475-4484.

[36] X. Fu, Y. Lin, D. M. Lin, D. Mechtersheimer, C. Wang, and F. Ameen, et al., "Bidcell: Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data," *Nature Communications*, vol. 15, no. 1, p. 509, 2024.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv Prep. arXiv:2010.11929,* 2021.

[38] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[39] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv Prep. arXiv:1803.08375*, 2018.

[40] B. R. Swain, K. J. Cheoi, and J. Ko, "SAM guided task-specific enhanced nuclei segmentation in digital pathology," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024.* Springer Nature, Switzerland, 2024.

[41] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv Prep. arXiv:1412.6980,* 2014.

[42] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, Ö. Smedby, and C. Wang,"A two-stage U-Net algorithm

for segmentation of nuclei in H&E-stained tissues," in *Digital Pathology: 15th European Congress, ECDP 2019,* Warwick, UK, 2019,pp. 75-82.

[43] B. R. Swain, K. J. Cheoi, and J. Ko, "Nuclei segmentation in histopathological images with enhanced U-Net3+," *Medical Imaging with Deep Learning,* 2024.

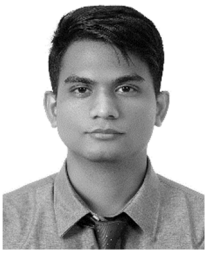
[44] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International, pp. 240-248, 2017.

[45] T. Y. Ross and G. K. H. P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Jul. 2017, pp. 2980-2988.

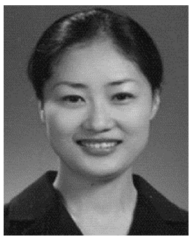
[46] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv Prep. arXiv:1505.00853*, 2015.

[47] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: A self-gated activation function," *arXiv Prep. arXiv: 1710.05941*, vol. 7, no. 1, p. 5, 2017.

AUTHORS



Bishal Ranjan Swain received his BS from College of Basic Sciences and Humanities, OUAT and M.S. degrees from the Department of Computer Engineering from Pondicherry University, India, in 2018 and 2020, respectively. In 2022, he joined the Department of Computer AI Convergence for pursuing his Ph.D. degree at Kumoh National Institute of Technology, Korea as a Global Korea Scholarship recipient. His research interests include computer vision, epecially material segmentation and medical segmentation, and in machine and deep learning.



Kyung Joo Cheoi received the B.S. degree in computer science from Chungbuk National University, Cheongju, Korea, in 1996, and the M.S. and Ph.D. degrees in computer science from Yonsei University, Seoul, Korea, in 2002. From 2002 to 2005, she worked as a Research Engineer with TI-Specialist-Technology of LG CNS, Seoul. Since 2005, she has been a Professor with the Computer Science Department, Chungbuk National University, Korea. Her research interests include computer vision, image processing, and machine learning.

Jaepil Ko received the B.S., M.S., and Ph.D. degrees from Yonsei University, Korea, in 1996, 1998, and 2004, respectively. Since 2004, he has been a Professor with the Computer Engineering Department, Kumoh National Institute of Technology, Korea. His research interests include computer vision, machine learning, and image processing.