Enhancing Cell Segmentation using Deep Learning Models by Custom Processing Techniques

Van-De NGUYEN*, Minh-Huong Hoang DANG†, Quang-Huy NGUYEN†, Manh-Cuong DINH †, Thanh-Ha DO †
* Department of Pathological Anatomy, Central Military Hospital 108
E-mail: doctorde@gmail.com
† VNU University of Science
E-mail: danghoangminhhuong_t65, nguyenquanghuy1_t66, dinhmanhcuong_t66, dothanhha (@hus.edu.vn)

***Abstract*—This study presents a comprehensive workflow for cell segmentation using the Monuseg dataset. Our method integrates several advanced techniques to enhance the performance of segmentation models. We apply Principal Component Analysis (PCA) and Gaussian Mixture Model - Expectation Maximization (GMM-EM) clustering for data pre-processing. The training and evaluation significantly improved the mean IoU metric, demonstrating the enhanced performance of Language meets Vision Transformer (LViT) compared to other deep-learning networks. This constructed workflow shows substantial potential in improving the accuracy of cell segmentation tasks, which is crucial and beneficial for various biomedical applications.**

I. Introduction

Cytopathology, the study of cellular-level diseases, is a pivotal tool in cancer screening. Artificial intelligence can potentially revolutionize diagnostics, especially segmenting cell images from micrographs.

Accurate cell segmentation is critical, the cornerstone of medical image analysis. It enables the precise identification and classification of cellular structures within microscopic images. It is significant for various biomedical applications, such as computational pathology, disease diagnosis, prognosis, and computer-aided diagnosis.

For the classical approach, the Split and Merge Watershed (SM-Watershed) method [1] for cell segmentation in fluorescence microscopy images. Initially, the Marker-Controlled Watershed (MC-Watershed) algorithm provides preliminary segmentation. The Split phase separates clusters using cell characteristics like size and convexity, while the Merge phase refines these segments by eliminating over-segmentation. This method effectively balances segmentation accuracy without needing labeled data.

The segmentation method that combines the K-L transform and the OTSU method is not just a standard cell segmentation method. It is a proven and effective technique. The K-L transform is utilized to select the most informative channel from the image, followed by the OTSU method to determine an optimal automatic threshold value for segmentation. This approach has been applied to various types of cell images, demonstrating its practicality and effectiveness.

While effective in specific scenarios, classical approaches such as Watershed, OTSU, and K-L transform often need help with the complexity and variability of all cell segmentation tasks. These methods can result in over-segmentation, under-segmentation, or mis-identification of cells. Consequently, recent research has increasingly focused on deep learning techniques, which offer improved accuracy and robustness in handling diverse and complex cell images (see detail in Section II).

Deep Learning-based innovative algorithms generally archive state-of-the-art performance in medical imaging segmentation. Among the models, we are inspired by the LViT (Language meets Vision Transformer) model [2] since it tackles the challenge of limited labeled data in medical image segmentation by incorporating medical text annotations. This text information supplements the image data, allowing the model to learn even with limited labeled examples.

We improve the performance of the LViT model by ensuring high-quality data input. In more detail, we integrate Principal Component Analysis (PCA) for dimensionality reduction and Gaussian Mixture Models with Expectation-Maximization (GMM-EM) in the pre-processing pipeline to optimize the clustering process. This has resulted in improved preliminary segmentation maps, a promising outcome that paves the way for further advancements in cell segmentation.

The following parts are organized as follows: Section II shows some base deep learning models applied in cell segmentation. Section III focuses on our contribution to the combination of GMM-EM and LVit model. Moreover, Section IV indicates the experiment results, and Section V determines this method's conclusion and discussion.

II. Deep Learning Approaches for Cell Segmentation in Microscopic Images

This section reviews some deep-learning approaches that have had good results in the last few years, particularly on the MoNuSeg dataset, a widely used benchmark dataset in cell segmentation. Understanding the performance of these approaches on this dataset will help us decide the direction of continuing research in this field.

In 2020, Jha *et al.* [3] proposed a Double U-Net model by adding another U-Net at the bottom of the network to capture supplementary semantic information efficiently. Further, Atrous Spatial Pyramid Pooling (ASPP) was adapted to capture contextual data, and the post-processing techniques significantly improved the result of automatic polyp detection. However, since this network uses two Unet models, the increase in the number of parameters is a limitation of this model.

Wang *et al.* [4] with a Bending Loss Regularized (BLR) model is successful in tackling the challenge of segmenting overlapped nuclei in histopathology images. This model applied high and low penalties to contour points with large and small curvature. In addition, the bending loss helps to avoid the generation of boundaries for two or more nuclei that are touching. The BLR model performs better than other models. However, there is still a problem with the segmentation of overlapping nuclei.

Hassan *et al.* [5] proposed a Pyramid Scene Parsing with SegNet (PSPSegNet) to identify and delineate the boundaries of nuclei. The experiment indicates that the PSPSegNet model is effective with F1-Score and AJI at 0.8815 and 0.7080, respectively. Concerning the object level, the PSPSegNet model relies on training data and, therefore, is unsuitable for exacting cell shapes.

The author Lagree *et al.* [6] proposed a gradient-boosting U-Net(GB U-Net) to segment breast tumor cell nuclei. This research shows that deep convolutional neural networks are suitable for training with transfer learning on a set of histopathological images independent of breast tissue to segment or not tumor nuclei of the breast.

In the same year, 2021, Li *et al.* [7] proposed the Bagging Ensemble Deep segmentation (BEDs) model, which aggregates self-ensemble learning and testing stage augmentation to improve the robustness of nucleus segmentation. However, this model needs to segment better when the images are complicated in shape and structure.

One year later, Qin *et al.* [8] proposed the REU-Net model to improve segmentation accuracy by focusing on region-specific features within images. It leverages enhanced feature extraction techniques to identify better and segment nuclei in medical images. However, while REU-Net boosts performance, it may introduce additional computational complexity due to its region-focused approach. Liang *et al.* [9] also use region information to build a model that integrates Guided Anchoring (GA) into the Region Proposal Network (RPN) and using a fusion box score (FBS) with soft non-maximum suppression (SoftNMS). This model improves accuracy over traditional CNN-based approaches and enhances cell-level analysis in digital tissue images.

In 2023, the large-scale model named Segment Anything Model (SAM) was introduced [10]. This model is trained on 11 million images with over 1 billion masks for general-purpose segmentation. Although SAM doesn't initially provide high-quality segmentation for medical images, its masks, features, and stability scores are valuable for improving medical image segmentation models. This model can be applied to augment inputs for models like U-Net, with experiments on three tasks that show its effectiveness.

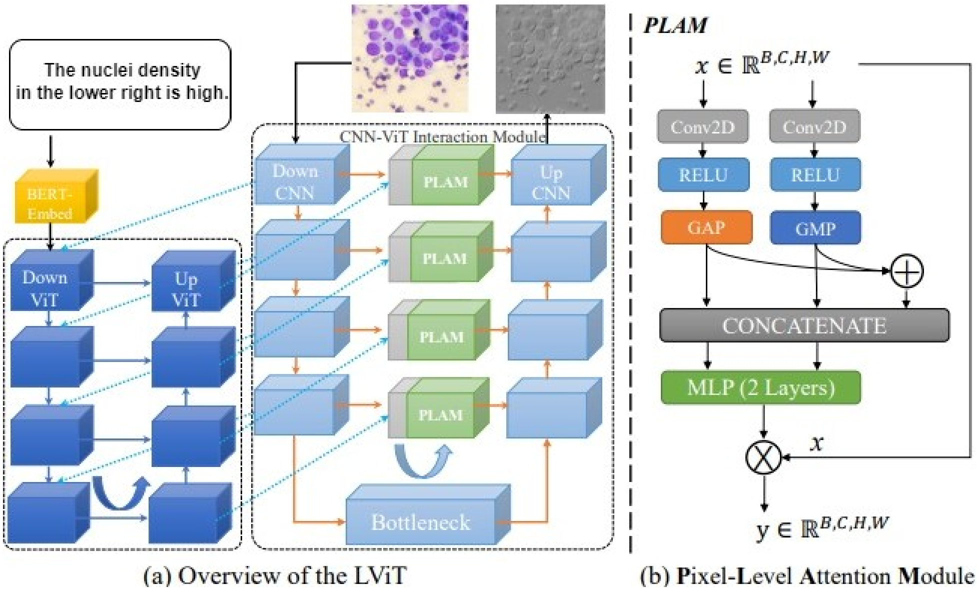
In recent years, U-Net and its variants have been widely used in pathology image segmentation, leveraging skip connections to recover detailed information. However, the semantic gap between encoder and decoder can hinder performance. To address this, the FusionU-Net [11] incorporates a fusion module to reduce semantic gaps by exchanging information between skip connections. Our two-round fusion design considers local relevance and bi-directional information exchange across layers. In addition, another model, named BiU-Net [12], combines CNNs and transformers using a two-stage fusion strategy. The Single-Scale Fusion (SSF) stage integrates local and long-range features, while the Multi-Scale Fusion (MSF) stage eliminates the semantic gap between deep and shallow layers. Additionally, a Context-Aware Block (CAB) in the bottleneck enhances multi-scale features in the decoder, improving segmentation performance.

III. The Proposed Approach: Combination LVit Model with GMM-EM Methods

Compared to the successful methods mentioned in Section II, the LViT (Language meets Vision Transformer) model [2] overcomes these limitations by significantly enhancing context awareness and reducing dependency on extensive labeled data, making it more suitable for cell segmentation tasks. LViT effectively captures global relationships within images, leading to more accurate and efficient segmentation results than Unet++. Thus, we enhanced this model in this paper by applying the pre-processing stage to improve performance (see Figure 2).

Before describing our contribution in more detail, we revise the main idea of the LViT model, as illustrated (Figure 1): LViT uses text information to supplement the image data, allowing the model to learn even with limited labeled examples. Furthermore, LViT leverages semi-supervised learning, utilizing text data to generate high-quality pseudo-labels for unlabeled images. An Exponential Pseudo Label Iteration (EPI) mechanism assists the Pixel-Level Attention Module (PLAM) preserve local image features during this process. The Exponential Pseudo Label Iteration (EPI) mechanism is a critical component in LViT, designed to enhance the quality of pseudo-labels for unlabeled image iterative. EPI refines the pseudo-labels in each iteration by incorporating feedback from the model's predictions, which become increasingly accurate over time. This iterative process improves pseudo-labels' reliability, facilitating more effective semi-supervised learning. Complementing EPI, the Pixel-Level Attention Module (PLAM) ensures that fine-grained details are preserved during segmentation. PLAM assigns attention scores to each pixel, allowing the model to focus on critical regions within the image and maintain local feature integrity. This dual mechanism of EPI and PLAM enables LViT to achieve high precision in cell segmentation tasks, even with limited labeled data.

Compared to models like U-Net++ and BiO-LinkNet, which rely solely on convolutional operations, LViT's transformer architecture allows it to capture local and global information


The nuclei density in the lower right is high.
BERT-Embed
Down ViT
Up ViT
CNN-ViT Interaction Module
Down CNN
PLAM
Up CNN
PLAM
PLAM
PLAM
Bottleneck
PLAM
$x \in \mathbb{R}^{B,C,H,W}$
Conv2D
Conv2D
RELU
RELU
GAP
GMP
CONCATENATE
MLP (2 Layers)
x
$y \in \mathbb{R}^{B,C,H,W}$

(a) Overview of the LViT

(b) Pixel-Level Attention Module

Fig. 1. Illustration of (a) the proposed LViT model and (b) the Pixel-Level Attention Module (PLAM). The proposed LViT model is a Double-U structure that combines a U-shape CNN branch with a U-shaped ViT branch [2].

more effectively. While MaxViT-UNet achieves high performance using a combination of convolutions and transformers, LViT's advantage lies in its ability to leverage image data and text annotations. This adaptability makes it more effective in cell segmentation tasks.

Our pre-processing stage prepares the data and ensures a robust and reliable process for LViT optimally. We use Principal Component Analysis (PCA) to reduce the data's dimensionality, followed by Gaussian Mixture Models (GMMs) with Expectation Maximization (EM) to group images into distinct clusters.

The Principle Component Analysis was first applied in our pre-processing pipeline to reduce the feature vectors' dimensions, including pixel intensity values and texture features extracted from the cell images. The PCA transformation, a crucial step in our pipeline, is defined in Equation 1.

$$\mathbf{z} = \mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}) \tag{1}$$

Where $\mathbf{x}$ is the original feature vector, $\boldsymbol{\mu}$ is the mean of the feature vectors, $\mathbf{W}$ is the matrix of eigenvectors of the covariance matrix, and $\mathbf{z}$ is the transformed feature vector in the principal component space.

Following dimensionality reduction, GMM-EM was applied for image clustering. GMM-EM was chosen due to its ability to effectively process diverse and complex cellular data, its flexibility in modeling complex distributions, and its capacity to provide soft probabilities for each cluster. The GMM model is defined as in Equation 2.

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

Where $\mathbf{z}$ is the PCA-transformed feature vector, K is the number of clusters, π_k is the weight of the k-th cluster, and $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density function of the Gaussian distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

An iterative experimental process meticulously determined the number of clusters K. This process was designed to optimize the separation between clusters and the compactness within each cluster, ensuring the highest quality results. Upon completion of the clustering process, each pixel in the image was labeled according to the cluster with the highest posterior probability, generating a preliminary segmentation map.

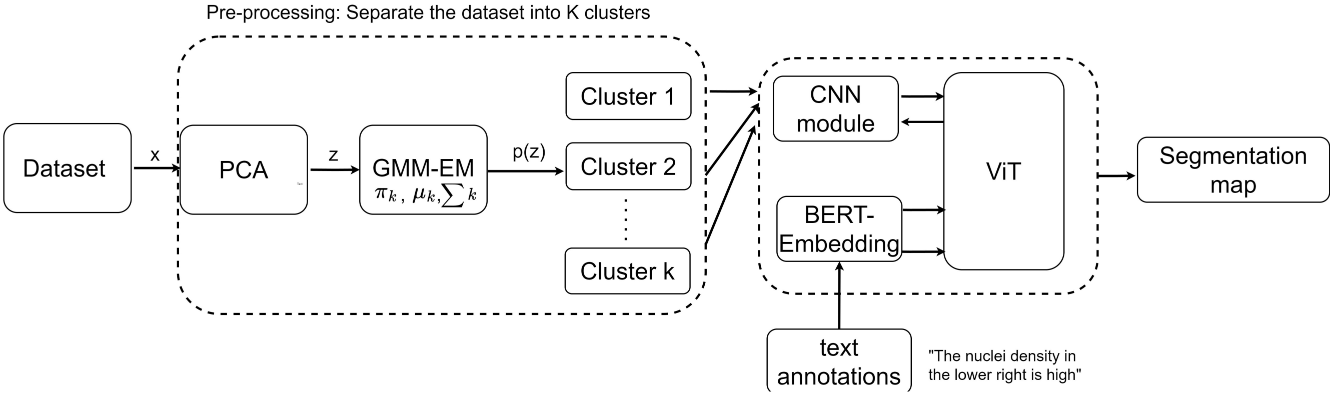
Our research has achieved several notable outcomes:

- Developed a pre-processing pipeline that effectively integrates PCA for dimensionality reduction with GMM-EM for clustering, providing high-quality input for the subsequent LViT model.
- Our research has led to a significant achievement-the combination of PCA and GMM-EM has optimized the clustering process. This has resulted in improved preliminary segmentation maps, a promising outcome that paves the way for further advancements in cell segmentation.

IV. Experiments Results

The experiment is evaluated using the highly regarded MICCAI MoNuSeg dataset. This dataset comprises 44 images, each sized at 1000×1000 pixels with $28,846$ labeled cell nuclei distributed across nine organs: breast, liver, kidney, prostate, bladder, colon, stomach, lungs, and brain. The dataset is organized into 24 images for training, 6 for validation, and 14 reserved for testing.

Our approach involved extensive experimentation with various sizes of image patches in images of the MoNuSeg2018



Pre-processing: Separate the dataset into K clusters
Dataset
x
PCA
z
GMM-EM
π_k , μ_k,$\sum k$
p(z)
Cluster 1
Cluster 2
Cluster k
CNN module
BERT-Embedding
ViT
Segmentation map
text annotations
"The nuclei density in the lower right is high"

Fig. 2. Proposed Model for Cell Segmentation over Microscopic Images combining GMM-EM to LViT Model

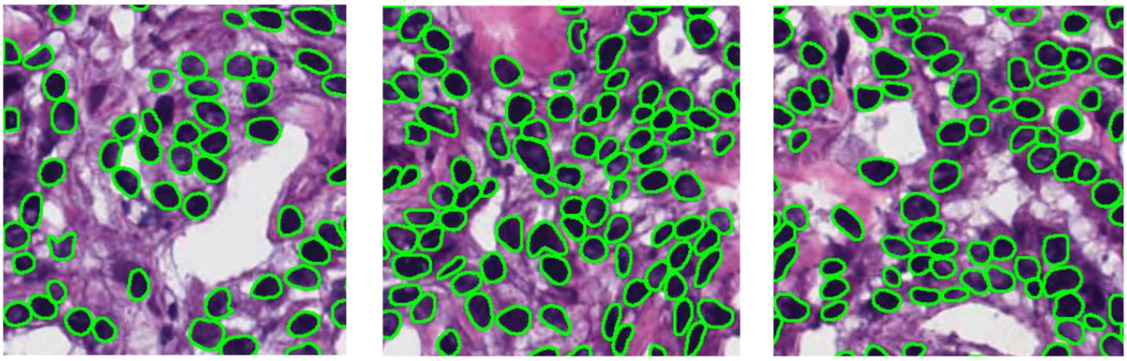


Fig. 3. Cell segmentation results on MoNuSeg2018 dataset [13] using the proposed model, achieving a Dice score of 0.80, IoU of 0.66, and a runtime of 17155.2 seconds on $K = 2$ clusters, trained on an NVIDIA A100 (Google Colab Pro).

Method	LVit [2]	Proposed model	MaxViT-UNet [14]	Dice Unet [15]	Unet++ [16]	BiO-LinkNet [17]	LinkNet [17]	R2U-Net [18]
Dice	0.78	**0.80**	0.83	0.76	0.77	0.77	0.77	0.80
IoU	0.65	**0.66**	0.72	0.62	0.63	0.62	0.63	0.68

TABLE I
SEMANTIC SEGMENTATION RESULTS ON THE MONUSEG2018 DATASET.

dataset. We were cropping images to 256×256 pixels produced trade-off results. We also addressed memory constraints by adopting overlapping patches with a 70-pixel overlap while maintaining the original organ distribution.

Given the relatively small dataset size of $1,100$ patches, this paper employs data augmentation techniques to enhance model training and accuracy, effectively expanding the dataset to $2,200$ images. These techniques, including random horizontal flipping, rotating, and adding a Gaussian filter with a random parameter, were developed to address the challenges of working with a small dataset, demonstrating our commitment to overcoming research obstacles.

The semantic segmentation model's accuracy is evaluated using the Intersection over Union (IoU) measure and the Dice coefficient. For the individual segmentation problem, the model's accuracy is assessed based on the Score value. The IoU, Dice, and Score measurements have values in the range [0, 1]; when the value is close to 0, the model's accuracy is low; the closer it is to 1, the higher its accuracy.

Using the LViT method, each image in the Monuseg2018 dataset is accompanied by a text passage providing a specific description and evaluation of that particular image. In more detail, each image describes the characteristics of the nuclei, such as evenly/sparse distribution and higher/lower density areas, etc. There may also be several images with the same text passage. We pre-processed the images using the GMM-EM (Gaussian Mixture Model with Expectation-Maximization) technique to effectively process the visual data before applying the LViT method. This process clusters only the image data, creating groups of visually similar images while maintaining the individual text descriptions for each image. We conducted experiments with different K values to determine the optimal number of clusters (K) for the image data. We adjusted based on experimental results and dataset characteristics to optimize model performance. This approach allows us to leverage both the clustered visual information and the unique textual descriptions for each image in our dataset, enhancing the overall performance of the LViT method.

K	Dice Score	IoU Score	Execution Time (seconds)
1	0.7659	0.6307	13449.25
2	**0.8015**	**0.6566**	17155.2
3	0.7039	0.5522	23643.5

TABLE II
PERFORMANCE METRICS WITH DIFFERENT NUMBERS OF CLUSTERS ON THE TEST DATASET.

Following the data pre-processing with GMM-EM, we experimented with different numbers of clusters K, starting from 1 and incrementally increasing. After training the LViT model from scratch on each clustered dataset, we selected $K = 2$ as it yielded the highest accuracy compared to other configurations. Choosing $K = 2$ allows the model to capture enough variance to segment the cells accurately while maintaining robustness against noise or unnecessary complexity. This balance likely leads to better generalization, as evidenced by the superior performance metrics. Table II shows the performance metrics for each K. The model's performance was then evaluated on the test set using metrics such as IoU and the Dice coefficient. The results (see Table I) are a testament to our progress, with the LViT method combined with GMM-EM achieving an IoU of 0.66 and a Dice coefficient of 0.8, surpassing the baseline LViT model (IoU = 0.65, Dice = 0.78). Compared with standard LViT and U-Net++, our proposed method demonstrates that our approach is competitive and a significant step forward in the microscopy image processing field.

Table I illustrates our results compared to another method. The results of the proposed method show significant improvement in cell segmentation performance on the MoNuSeg2018 dataset compared to other methods. Our approach achieved the second highest Dice score of 0.80 and an IoU of 0.66, demonstrating better accuracy and reliability. In contrast, the standard LViT method scored a Dice of 0.78 and IoU of 0.65, indicating a noticeable enhancement with improvements. The proposed method outperforms both metrics compared to Dice Unet and Unet++, which scored Dice values of 0.76 and 0.77 and IoU values of 0.62 and 0.63, respectively.

Although MaxViT-UNet achieves a higher Dice score of 0.83, its superior performance can be attributed to its multi-axis attention mechanism, which better captures local and global contexts. This, together with combining convolutional layers and transformers, enhances its segmentation accuracy. However, its increased complexity and computational demands highlight the trade-offs, whereas the proposed method offers a strong balance between performance and efficiency.

Other methods, like BiO-LinkNet and LinkNet, with Dice scores of 0.77 and IoU scores of 0.62 and 0.63, respectively, also fall short of our results. The R2U-Net method also performs well, with a Dice score of 0.80 and an IoU of 0.68. However, our method remains robust and consistent across both evaluation metrics, reinforcing its effectiveness for accurate and efficient cell segmentation.

Figure 3 presents three cropped images from the original MoNuSeg2018 dataset. The figure indicates that using the GMM-EM and LViT effectively segments overlapping cells and distinguishes between cells and other components, such as blood, which are not the focus of this study.

V. CONCLUSIONS AND DISCUSSIONS

This paper presented an enhanced cell segmentation approach by integrating the LViT model with GMM-EM preprocessing. The use of clustering is crucial in improving the learning process, as it allows the model to capture more distinct and characteristic features from the data. By applying PCA for dimensionality reduction, followed by GMM-EM clustering, the data is effectively segmented into meaningful groups, making it easier for the model to focus on relevant patterns. This structured input significantly boosts the accuracy and reliability of the model. This combination achieved superior performance when evaluating the MICCAI MoNuSeg dataset. In the future, we plan to collect data from diverse sources and further optimize the model parameters to enhance the accuracy of our approach significantly.

ACKNOWLEDGMENT

This research was funded by the research project QG.23.71 of Vietnam National University, Hanoi.

REFERENCES

[1] M. Gamarra, E. Zurek, H. J. Escalante, L. Hurtado, and H. San-Juan-Vergara, "Split and merge watershed: A two-step method for cell segmentation in fluorescence microscopy images," *Biomedical signal processing and control*, vol. 53, p. 101 575, 2019.

[2] Z. Li, Y. Li, Q. Li, *et al.*, "Lvit: Language meets vision transformer in medical image segmentation," *IEEE transactions on medical imaging*, 2023.

[3] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, *Doubleu-net: A deep convolutional neural network for medical image segmentation*, 2020. arXiv: 2006.04868 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2006.04868.

[4] H. Wang, M. Xian, and A. Vakanski, *Bending loss regularized network for nuclei segmentation in histopathology images*, 2020. arXiv: 2002.01020 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2002.01020.

[5] L. Hassan, A. Saleh, M. Abdel-Nasser, O. A. Omer, and D. Puig, "Promising deep semantic nuclei segmentation models for multi-institutional histopathology images of different organs," 2021.

[6] A. Lagree, M. Mohebpour, N. Meti, *et al.*, "A review and comparison of breast tumor cell nuclei segmentation performances using deep convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, p. 8025, 2021.

[7] X. Li, H. Yang, J. He, *et al.*, *Beds: Bagging ensemble deep segmentation for nucleus segmentation with testing stage stain augmentation*, 2021. arXiv: 2102.08990 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2102.08990.

[8] J. Qin, Y. He, Y. Zhou, J. Zhao, and B. Ding, "Reu-net: Region-enhanced nuclei segmentation network," *Computers in Biology and Medicine*, vol. 146, p. 105 546, 2022.

[9] H. Liang, Z. Cheng, H. Zhong, A. Qu, and L. Chen, "A region-based convolutional network for nuclei detection and segmentation in microscopy images," *Biomedical Signal Processing and Control*, vol. 71, p. 103 276, 2022.

[10] Y. Zhang, T. Zhou, S. Wang, P. Liang, Y. Zhang, and D. Z. Chen, "Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 129–139.

[11] Z. Li, H. Lyu, and J. Wang, "Fusionu-net: U-net with enhanced skip connection for pathology image segmentation," in *Asian Conference on Machine Learning*, PMLR, 2024, pp. 694–706.

[12] Z. Huang, Y. Zhao, Z. Yu, *et al.*, "Biu-net: A dual-branch structure based on two-stage fusion strategy for biomedical image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 252, p. 108 235, 2024.

[13] A. Goodman, A. Carpenter, E. Park, *et al.*, *2018 data science bowl*, Kaggle Competition, 2018. [Online]. Available: https://kaggle.com/competitions/data-science-bowl-2018.

[14] A. R. Khan and A. Khan, "Maxvit-unet: Multi-axis attention for medical image segmentation," *arXiv preprint arXiv:2305.08396*, 2023.

[15] K. C. T. Nguyen, "Segment automatically cells over microscopic images based on deep learning techniques," *Undergraduate thesis in VNU University of Science*, 2021.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and 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*, Springer, 2018, pp. 3–11.

[17] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*, IEEE, 2017, pp. 1–4.

[18] M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, "Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net)," in *NAECON 2018-IEEE National Aerospace and Electronics Conference*, IEEE, 2018, pp. 228–233.