NuSegFormer: Enhanced Cell Nuclei Segmentation Using a Hybrid UNet–Vision Transformer with Gated Multi-Head Attention

Aditi Damodar
Department of Artificial Intelligence and Data Science
Koneru Lakshmaiah Education Foundation
Hyderabad, India
2310080036@klh.edu.in

Hitesh Tekchandani
Department of Artificial Intelligence and Data Science
Koneru Lakshmaiah Education Foundation
Hyderabad, India
hitesh.t@klh.edu.in

Praveen Kumar Shukla
School of Computing and Intelligent Systems
Manipal University Jaipur
Jaipur, India
praveen.shukla@jaipur.manipal.edu

Anish Kumar Vishwakarma
School of Computer Science
UPES Dehradun
Dehradun, India
anishk.vishwakarma@ddn.upes.ac.in

***Abstract*—This paper introduces a NuSegFormer for segmenting cell nuclei, focusing on creating accurate segmentation masks by effectively distinguishing the foreground from the background. We employed a Hybrid UNet-ViT architecture that incorporates Gated Attention Mechanism and ReGLU activation function to enhance feature significance and boundary detection. The five-gate attention mechanism allows precise regulation of information flow at each phase of attention computation, while the ReGLU activation selectively processes features through a trained gating stream, collectively boosting the model's discriminative power. Experimental findings reveal that our approach achieves a Dice Score of 93.0% and an IoU of 86.92%, confirming its efficacy in cell nuclei segmentation tasks.**

***Index Terms*—Nuclei Segmentation, Gated Attention Mechanism, ReGLU Activation, ViT, UNet**

I. Introduction

Segmenting medical images is crucial for disease diagnosis and treatment, yet accurately segmenting cell nuclei is difficult due to variations in their size and shape, overlapping clusters, and image noise. The cell nucleus acts as the control center for most human cells, and changes in its size and structure are key indicators of cancer, tumors, and genetic disorders, making precise segmentation of nuclei essential in histopathology and medical imaging research. However, current deep learning models often struggle to capture long-range dependencies and have difficulty with densely overlapping clusters, which limits their segmentation accuracy.

The primary objective of this study is to design an improved Vision Transformer-based framework featuring a Gated Attention mechanism that greatly enhances the performance of nucleus segmentation. The authors introduce NuSegFormer a Hybrid UNet-ViT incorporating Gated Multi-Head Attention [1] and ReGLU Activations [2]. In this model, a pretrained EfficientNetV2S [3] encoder captures local texture and edge details across various scales, while the Vision Transformer bottleneck addresses long-range spatial relationships. The five-gate attention mechanism allows for precise regulation of information flow at each stage of attention computation, filtering out irrelevant features and emphasizing important ones, which leads to more accurate and reliable nucleus segmentation on the 2018 DSB dataset [4].

Section II outlines the related existing work, section III illustrates the methodology, section IV describes the dataset and the experimentation setup, section V gives details about the results, and section VI concludes.

II. Literature Review

Gautam et al. [5] implemented a U-Net CNN architecture for nuclei segmentation and evaluated it on the 2018 DSB dataset. Imtiaz et al. [6] proposed ConDANet, a contourlet-driven attention network combining content-preserving sampling, edge-guided control, contourlet features, and wavelet pooling. Alom et al. [7] proposed multiple deep learning models for nuclei analysis, including DCRN for classification, R2U-Net for segmentation, and UD-Net for detection, all evaluated on the 2018 nuclei challenge dataset. Tran et al. [8] introduced Trans2UNet, a modified U-Net for spatial refinement with a transformer-based patch encoder and a lightweight WASP-KC module. Pham et al. [9] proposed seUNet-Trans, a UNet-Transformer hybrid integrating a transformer bridge through pixel-wise embedding and spatially reduced attention. Ramya et al. [10] introduced a Transformer-based encoder integrated with a U-Net architecture to improve nuclei segmentation accuracy.

While the above studies have investigated segmentation on the 2018 DSB dataset [4], many approaches still fall short in robustness and effectiveness for real-world deployment. To ad-

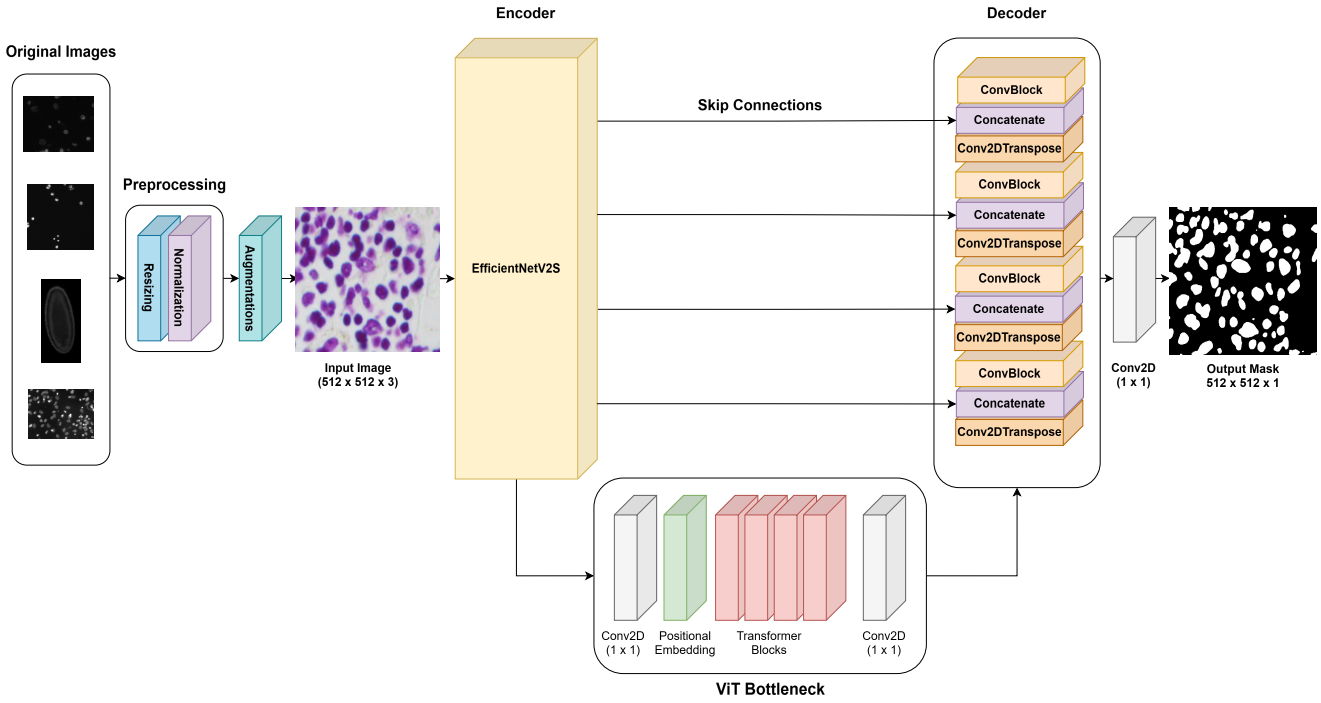


Fig. 1: Workflow of the proposed NuSegFormer

dress this, the authors propose a Hybrid UNet-ViT with Gated Multi-Head Attention and ReGLU Activation. Five learnable sigmoid gates control the flow of query, key, value, attention output, and final output, suppressing irrelevant features while amplifying discriminative ones. The ReGLU activation further strengthens representation by splitting projections into content and gate streams, enabling selective feature propagation and producing sparser, more informative activations compared to standard GELU activations.

III. Methodology

A. Details of Architecture

The authors propose a Hybrid UNet-ViT model incorporating a Gated Multi-Head Attention Mechanism, whose complete workflow is illustrated in Fig. 1. As shown in Fig. 2, EfficientNetV2-S [3] replaces the standard U-Net's encoder, serving as the backbone for feature extraction to obtain multi-scale feature maps. It incorporates skip connections at four different scales and includes a bottleneck feature map produced at 16×16×1280. These feature maps are passed to the ViT Bottleneck, where a 1×1 convolution projects the 1280 channels down to 256, and the spatial grid is reshaped into a sequence of 256 tokens.

To boost the network's capacity for capturing global contextual information and long-range spatial dependencies, a Vision Transformer (ViT) is integrated at the bottleneck layer. This layer deals with a highly compressed feature representation that has reduced spatial dimensions, which significantly decreases the computational burden of self-attention. Consequently, the model can effectively gather information from distant parts of the image, resulting in more enriched semantic feature representations and enhanced segmentation accuracy. This is particularly beneficial for structures with intricate shapes, varying sizes, or unclear boundaries. Learned positional embeddings are then added, and the tokens are processed through four Transformer Blocks, each comprising a Gated Multi-Head Attention layer with 8 heads and 5 sigmoid gates, followed by a ReGLU activation. The tokens are subsequently reshaped back into a 16×16×256 spatial grid and passed through a final 1×1 convolution for channel mixing. The bottleneck output is then fed into the UNet decoder, which progressively upsamples through four stages, concatenating the corresponding encoder skip connection at each scale to recover fine-grained spatial detail. A final transposed convolution restores the full 512×512 resolution, and a 1×1 convolution with sigmoid activation produces the per-pixel segmentation mask.

B. Gated Attention

Gated Attention [1] is a combination of the standard multi-head attention and the gating mechanism.

(I) Standard Multi-Head Attention:

In Standard Multi-Head Attention, the input is transformed into queries, keys, and values, which are then used to calculate attention scores and generate an output. However, there is no regulation over the distribution of input into Q, K, V, or the extent of output utilization.

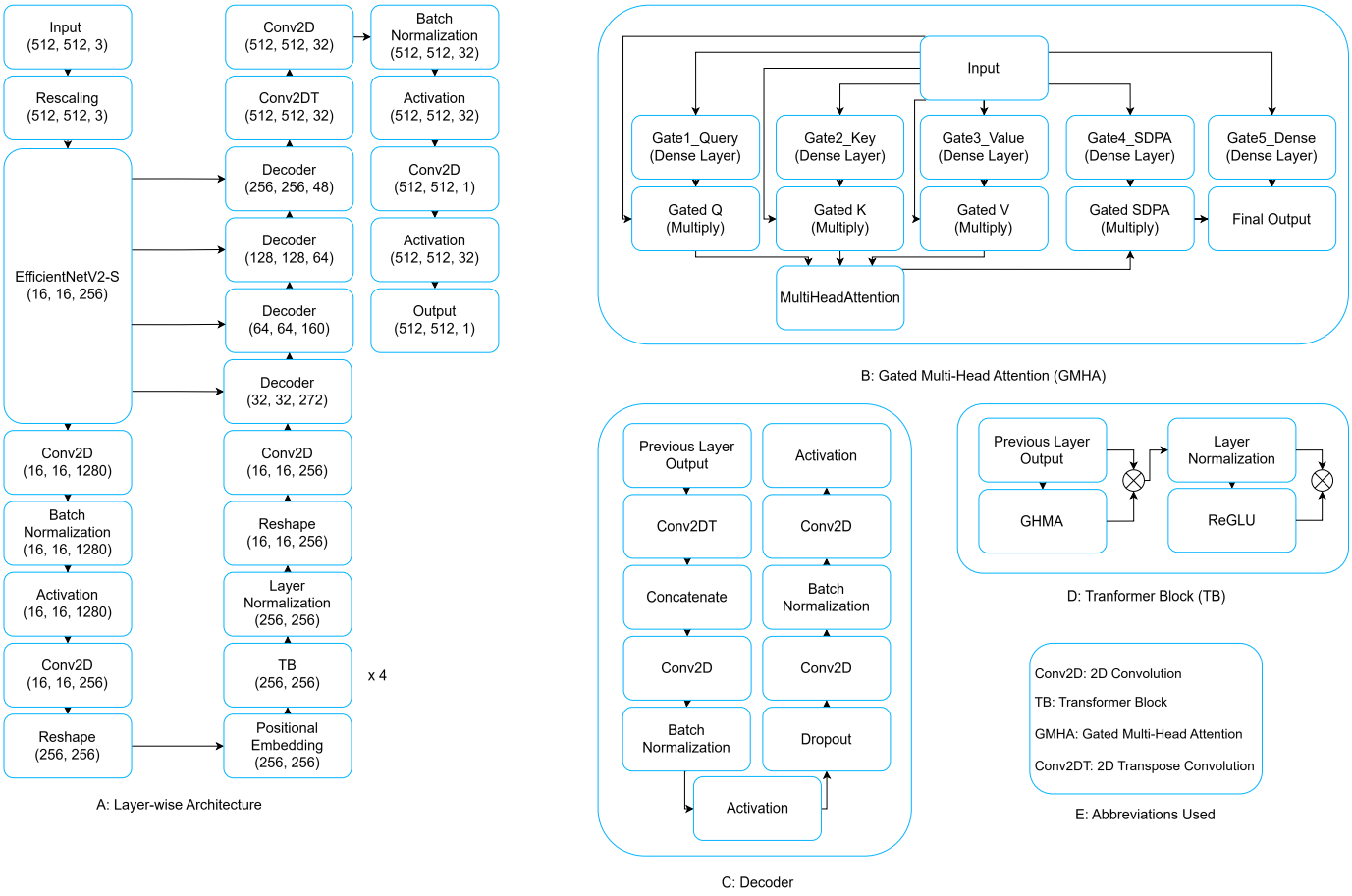


Fig. 2: Layer-wise architecture of the proposed NuSegFormer

(II) Gating Mechanism:
The gating mechanism resolves this by incorporating a sigmoid gate that generates values ranging from 0 to 1. By multiplying any tensor with a sigmoid gate, the model gains learned control over the flow of information.
This gating is implemented at five distinct stages within the attention mechanism.
(i) Gate 1 - Query Gate: Screens the query prior to attention, determining what each token can search for in other tokens.
(ii) Gate 2 - Key Gate: Screens the key before attention, managing what information each token shares with others.
(iii) Gate 3 - Value Gate: Screens the value before attention, controlling the content retrieved when a token is attended to.
(iv) Gate 4 - SDPA Gate: Filters the raw attention output using the original input, deciding how much of the attention result to retain.
(v) Gate 5 - Dense Gate: The final filter applied on top of the SDPA-gated output, again based on the original input, offering a last chance to suppress or enhance information before the residual connection.

C. ReGLU Activation

ReGLU [2] divides the projection into two separate channels - one for content and the other for a gate - and then combines them by multiplication (Equation (1)). The gate channel determines which elements of the content channel are allowed to pass through (Equation (2)).

$$\text{ReGLU}(x) = x_1 \odot \text{ReLU}(x_2) \tag{1}$$

$$y = (x_1 \odot \text{ReLU}(x_2))W_{\text{down}} + b \tag{2}$$

Here:

- xW_{up} projects the input to $2 \times$ MLP_DIM dimensions.
- The resulting vector is split into two equal halves: x_1 and x_2.
- x_1 is the **value (content)** channel that carries the information.
- x_2 is the **gate** channel that controls what passes through.
- $\odot$ denotes element-wise multiplication.
- W_{down} is the down-projection weight matrix of the feed-forward block.

IV. Dataset Details and Experimentation Setup

A. Dataset Details

The 2018 Data Science Bowl dataset [4] was used in this research, which includes histopathological and microscopy images. The initial dataset consists of 670 images with different dimensions.

B. Preprocessing and Augmentations

Images were resized to 512×512 and normalized to a [0, 1] range, while multiple masks per image were merged and converted into a single binary mask. To improve model learning, real-time augmentations - including horizontal and vertical flips, Gaussian blur ($\sigma = 2$ and $\sigma = 4$), and rotations ($\pm 270^{\circ}$) - were applied to the training set of 536 images, with the augmented and original images combined to form the final training dataset of 3216 images.

C. Training Configuration

The model was trained with a batch size of 8 using the AdamW optimizer (weight decay of 1×10^{-4}) and an initial learning rate of 3×10^{-4} with cosine decay and warm restarts. The AdamW optimizer enhances model generalization and minimizes overfitting by separating weight decay from adaptive gradient updates, resulting in more stable and efficient training. A hybrid Binary Cross-Entropy and Dice loss function was employed alongside five-fold cross-validation over 40 epochs, with all experiments conducted on an NVIDIA A100 GPU in Google Colab.

V. Results

A. Quantitative Results

Table I reports the quantitative performance of the proposed model. Two metrics: Dice Score (%), IoU (%) were used for the evaluation of the proposed model.

TABLE I: Experimentation Results

Fold	Dice Score (%)	IoU (%)
1	93.11	87.11
2	92.85	86.66
3	91.42	84.2
4	92.44	85.94
5	92.28	85.67
Mean ± Std	0.9242 ± 0.0058	0.8592 ± 0.0100
Highest	**93.00**	**86.92**

Table II summarizes the results of all the state-of-the-art architectures explored in this study, comparing them based on the average Dice Score (%).

TABLE II: Comparison with SOTA Architectures

Model	Avg. Dice Score (%)
R2UNet	90.94
Attention ResUNet	90.66
ResUnet	87.48
UNet	90.276
UNet++	89.88
Yolo	88.43
TransUNet	92.4
Hybrid + Gated (Proposed)	**93.00**

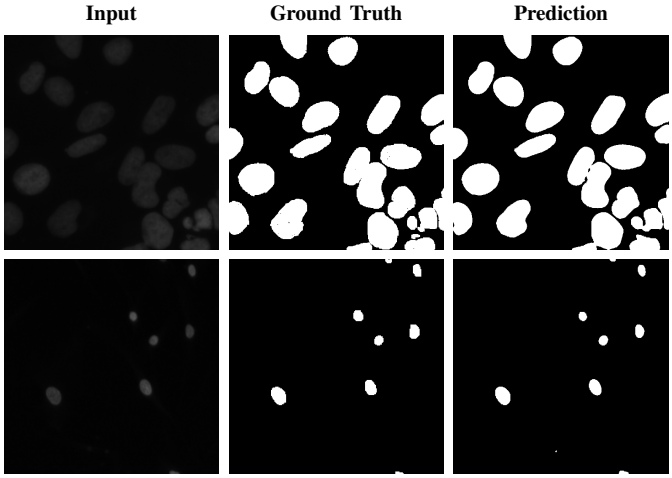
The proposed model achieved an average inference time of 143.09 ms per image.
As shown in Table III, the proposed approach outperforms existing methods with respect to Dice Score and IoU.

TABLE III: Comparison of Existing Methods

Authors	Methodology	Dice Score (%)	IoU (%)
Gautam et al. [5]	U-Net	-	88
Imitiaz et al. [6]	ConDANet	88.91	80.81
Tran et al. [8]	Trans2UNet	92.25	86.13
Pham et al. [9]	seUNetTrans	92.8	86.7
Ramya et al. [10]	TR-UNet	92.6	86.22
Proposed	**Hybrid + Gated**	**93.00**	**86.92**

Table IV presents the qualitative results, illustrating the model's predicted segmentation masks alongside the corresponding ground truth for a representative set of test images.

TABLE IV: Visualization of Input and Predictions



B. Ablation Studies

An ablation study was performed using three model variants:
(i) U-Net + ViT + 5-Gated Attention
(ii) EfficientNetV2-S + ViT + U-Net Decoder
(iii) EfficientNetV2-S + ViT + U-Net Decoder + 5-Gated Attention (proposed).
The results are presented in Table V.

TABLE V: Ablation Study

Architecture	Dice Score (%)
U-Net + ViT + 5-GA	91.75
EfficientNetV2-S + ViT + U-Net Decoder	92.68
EfficientNetV2-S + ViT + U-Net Decoder + 5-GA	93.00

Note: 5-GA = 5-Gated Attention

As shown in Table V, replacing the conventional U-Net encoder with EfficientNetV2-S increased the Dice Score from 91.75% to 92.68%, demonstrating the effectiveness of enhanced multi-scale feature extraction. Furthermore, incorporating the proposed 5-Gated Attention mechanism improved the Dice Score from 92.68% to 93.00%, indicating that Gated Attention facilitates more effective contextual feature

modeling within the ViT bottleneck. Overall, the complete architecture achieved a Dice Score of 93.00%, representing a 1.25 percentage-point improvement over the baseline model.

VI. Conclusion

This study introduces NuSegFormer, a Hybrid UNet and ViT model with Gated Attention and ReGLU activation, improving segmentation performance with higher Dice Score and IoU. The integration enhances feature learning and demonstrates strong potential for medical image segmentation. Future work will explore lightweight transformer variants and multi class tasks. Additionally, using larger datasets and diverse imaging modalities can improve generalizability and clinical applicability.

References

[1] Z. Qiu *et al.*, "Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free," 2025.

[2] N. Shazeer, "Glu variants improve transformer," 2020.

[3] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, 2021, pp. 10 096–10 106.

[4] A. Goodman *et al.*, "2018 data science bowl," https://kaggle.com/competitions/data-science-bowl-2018, 2018, kaggle Competition Dataset.

[5] B. Gautam and P. Gautam, "Automated nuclei detection in microscopy images," *International Journal on Engineering Technology*, vol. 2, no. 2, pp. 77–89, 2025.

[6] T. Imtiaz, S. A. Fattah, and M. Saquib, "Condanet: Contourlet driven attention network for automatic nuclei segmentation in histopathology images," *IEEE Access*, vol. 11, pp. 129 321–129 330, 2023.

[7] Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, "Microscopic nuclei classification, segmentation and detection with improved deep convolutional neural network approaches," Tech. Rep., 2021.

[8] D.-P. Tran, Q.-A. Nguyen, V.-T. Pham, and T.-T. Tran, "Trans2unet: Neural fusion for nuclei semantic segmentation," http://arxiv.org/abs/2407.17181, 2024, arXiv preprint.

[9] T. H. Pham, X. Li, and K. D. Nguyen, "Seunet-trans: A simple yet effective unet-transformer model for medical image segmentation," *IEEE Access*, vol. 12, pp. 122 139–122 154, 2024.

[10] H. P. R. Shree, Minavathi, and M. S. Dinesh, "Transformer based framework for nuclei segmentation and cell counting on microscopic images of multiple organs," *International Journal of Intelligent Networks*, vol. 6, pp. 283–292, 2025.