ISSN: 3067-2317
Research Article

World Journal of Artificial Intelligence and Robotics Research

Exploring Architectural Enhancements for Hover-Net: Deep Learning-Based Method of Automated Nuclear Segmentation and Classification

Shizhuo Qu*

Department of Computer Science, University of Toronto Scarborough

***Corresponding author:** Shizhuo Qu, Department of Computer Science, University of Toronto Scarborough.

Submitted: 18 July 2025 **Accepted:** 25 July 2025 **Published:** 31 July 2025

https://doi.org/10.63620/MK.WJAIRR.2025.1018

Citation: *Shizhuo Qu .(2025). Exploring Architectural Enhancements for Hover-Net: Deep Learning-Based Method of Automated Nuclear Segmentation and Classification. Wor Jour of Arti inte and Rob Res, 2(4), 01-10.*

Abstract

Accurate nuclei segmentation and classification are vital for computational pathology. This study builds upon the HoVer-Net architecture by integrating modern architectural components to enhance multi-task performance on the PanNuke dataset, which contains both segmentation and classification labels across 19 tissue types. We evaluate the effects of Squeeze-and-Excitation (SE) blocks, multi-head attention, enhanced DenseBlock decoders, and transformer-based encoders (ViT, SwinViT). All models follow HoVer-Net's preprocessing, training, and loss functions for consistent comparison. Results show that adding SE blocks to the encoder improves overall performance by approximately 3.6% in Dice scores, while transformer-based encoders lead to slight performance degradation. Our best model, MSDHV-Net (Multi-head Attention + SE + enhanced decoder), consistently outperforms the original HoVer-Net across several nuclei classes without increasing computational cost. These findings highlight the value of targeted architectural enhancements in advancing nuclei analysis models.

Keywords: Nuclear Instance Segmentation, Histopathology Image Analysis, Multi-task Learning, Vision Transformer (ViT), Swin Transformer, Squeeze-and-Excitation (SE) Module, Deep Learning, Hover Net.

Introduction

Nuclear instance segmentation is a fundamental task in computational pathology, underpinning a range of downstream applications such as cell classification, tissue phenotyping, and tumor microen- vironment analysis [1]. Accurate delineation of individual nuclei is critical for automated diagnosis and quantitative analysis in histopathological images, which are often challenged by dense cellular arrangements, overlapping nuclei, and diverse morphological variations [2].

Conventional image processing techniques have shown limited robustness in these complex settings due to their sensitivity to occlusion and shape variability. In contrast, deep learning approaches-particularly convolutional neural network CNN-based models like U-Net and nnU- Net-have greatly advanced performance by learning semantic features directly from data [3, 4]. Among these, HoVer-Net has emerged as a state-of-the-art multi-task architecture capable of simultane- ously predicting nuclear pixel maps (NP), horizontal–vertical distance maps (HoVer), and nuclear type maps (TP). This design has established a strong benchmark for nuclear instance segmentation and classification, particularly on large-scale datasets such as PanNuke [5, 6].

However, despite its effectiveness, HoVer-Net is built upon a residual CNN encoder, which may constrain its ability to capture long-range dependencies and global context. Recent developments in Transformer-based models—such as the Vision Transformer (ViT), Swin Transformer, and hybrid models like CellViT —have demonstrated powerful capabilities in modeling spatial relationships, which are especially important for the heterogeneous patterns seen in histopathology images [7-10]. Nevertheless, the incorporation of Transformer backbones into multi-task segmentation frameworks like HoVer-Net remains relatively underexplored.

In this study, I present a systematic investigation into enhancing the HoVer-Net architecture through the integration of advanced modules and modern design principles. My key contributions

are as follows:

1. I investigate the integration of Squeeze-and-Excitation (SE) blocks into the original CNN-based encoder of HoVerNet, aiming to enhance channel-wise feature recalibration [11]. This modifica- tion leads to measurable improvements in segmentation accuracy.
2. I explore the replacement of Hover Net's residual CNN encoder with Transformer-based modules, including Vision Transformer (ViT) and Swin Transformer. A unified architecture is proposed to preserve compatibility with HoVerNet's three-branch output structure.
3. I design an enhanced decoder architecture that leverages dense connections and dropout regular- ization, facilitating improved information flow and mitigating overfitting risks during training.
4. I conduct comparative experiments against the current state-of-the-art model, CellViT, and perform ablation studies to assess the individual contributions of each architectural component.

All experiments are conducted on the PanNuke dataset, following consistent training proto- cols-such as fixed epoch count, learning rate, pretrained models, optimizer, and train-validation splits-to ensure fair evaluation. While not all proposed models surpass the original HoVerNet in every metric, the findings emphasize the value of attention mechanisms and architectural enhancements in advancing segmentation and classification performance.

Materials and Methods

Dataset

All experiments in this study were conducted on the publicly available PanNuke dataset, a large-scale benchmark specifically curated for nuclei instance segmentation and classification tasks. PanNuke contains pixel-wise annotated histopathology image patches spanning 19 distinct tissue types. Each nucleus within the dataset is labeled according to one of five predefined categories: neoplastic, inflammatory, connective, dead, and epithelial.

The dataset was chosen for its comprehensive and high-quality annotations, as well as its estab-lished usage in prior research. These characteristics make PanNuke particularly suitable for evaluating multi-class nuclear segmentation methods and enable consistent comparisons with other state-of-the-art approaches.

Preprocessing and Augmentation

To ensure fair and consistent evaluation across all models, I adopted the preprocessing and augmentation procedures from the original HoVerNet pipeline. Input histopathology images were first normalized and resized, followed by padding to align with architectural constraints and to maintain spatial consistency across samples. Random shuffling of training data was applied to mitigate sampling bias and enhance the models' generalization performance[12].

The Pan Nuke dataset was partitioned into training, validation, and testing subsets using a ratio of 2:6:2, ensuring a balanced distribution of tissue types and cell classes across splits.

Model Variants Based on HoVerNet Architecture

To investigate the effects of architectural modifications on nuclei instance segmentation and classification, we developed six variants of the original HoVerNet framework. Each variant retains HoVerNet's distinctive three-head output structure-np_map, hv_map, and tp_map-which enables joint nuclear segmentation and classification. The modifications focus on enhancing the encoder and decoder designs, and are detailed as follows (the pretrained model can be found in Appendix A and training code in Appendix B):

(1) HoVerNet + SE (HoverNetEnhanced): This variant augments the encoder with Squeeze-and-Excitation (SE) blocks, which are integrated within the residual units. SE blocks perform adaptive channel-wise recalibration, emphasizing informative features while suppressing less useful ones, thereby enhancing representational capacity.

(2) HoVerNet + Multi-head Attention (Multihead-HoverNet): In this model, multi-head self- attention (MHSA) modules are embedded into the encoder to capture long-range dependen- cies and global contextual cues [13]. Inspired by prior analysis of attention heads, this design seeks to improve performance on complex spatial structures [14].

(3) HoVerNet + SE + MHSA + Enhanced Decoder (MS-DHV-Net): This is the most comprehen-sive CNN-based modification. It combines both SE and MHSA modules in the encoder and introduces a redesigned decoder featuring deeper DenseBlock structures, additional skip con-nections, and convolutional refinement layers. This design aims to facilitate robust feature propagation and enhanced spatial resolution restoration [15].

(4) HoVerNet + ViT Encoder (HoverViTNet): Here, the conventional CNN encoder is replaced with a Vision Transformer (ViT)[7], enabling the model to extract patch-wise global representations using self-attention. These transformer-derived features are decoded via a CNN-based decoder, allowing comparative evaluation of attention-driven global context modeling [16].

(5) HoVerNet + Custom SwinViT Encoder (HoVerIT): This architecture integrates a custom Swin Transformer encoder into the HoVerNet pipeline. The hierarchical design and shifted window self-attention in SwinViT capture both local and global dependencies more effectively than vanilla ViT [17]. The idea draws inspiration from the Swin-UNETR architecture, while maintaining compatibility with HoVerNet's three-branch outputs [18].

(6) HoVerNet + SwinUNETR from MONAI (HoverSwinNet): In this variant, the encoder is directly replaced with the SwinUNETR backbone from MONAI. Transformer-derived multi-scale features are routed through HoVerNet's original three-branch decoders, serving as a strong baseline to assess the integration feasibility and performance of prebuilt transformer encoders.

Training Configuration

All models were trained using the Adam optimizer with a fixed learning rate of 1e-4. Pre- trained ResNet-50 weights were used to initialize the encoder when applicable [19, 20]. Although alterna- tive optimizers such as AdamW were evaluated, they did not produce noticeable improvements in performance [21]. Attempts to modify the learning rate led to unstable training dynamics, including instances of gradient explosion, thereby justifying the choice of a conservative fixed schedule. Each model was trained for 80 epochs, based on empirical observations that validation perfor-mance typically saturated between epochs 50 and 60(See Figure A1). A batch size of 16 was used to optimize memory utilization on an NVIDIA L20 GPU with 44 GB

VRAM.

The loss functions followed the original HoVerNet formulation: binary cross-entropy (BCE) combined with Dice loss for nuclear pixel (np_map) and nuclear type (tp_map) predictions, and a joint loss of mean squared error (MSE) and mean squared gradient error (MSGE) for the horizontal-vertical distance (hv_map) regression task.

Evaluation Metrics

To comprehensively evaluate model performance, we employed several quantitative metrics: DICE coefficient, Panoptic Quality (PQ), Precision, F1-score, and Recall. These metrics were com-puted separately for each predicted cell type, enabling a fine-grained analysis of segmentation and classification performance. This multi-metric evaluation framework facilitated rigorous comparisons across different architectural variants and offered insight into the effectiveness and limitations of each proposed enhancement [22].

Experiment

To evaluate the effectiveness of our proposed methods, we conduct extensive experiments on
the Pan Nuke dataset, which contains nuclei annotations across multiple tissue types. All models are
trained and validated using the same data preprocessing pipeline and patch-based augmentation
strategy as described in the Methods section.

For the pathological modality, we fine-tune CellViT with customized architectural enhancements such as SE blocks, using a combination of focal, dice, BCE, and regression-based losses tailored to multi-task segmentation. Training is performed on 256×256 image tiles with overlapping strides, using the Adam optimizer with an initial learning rate of 1e-4 for 80 epochs. Pretrained weights from SAM-ViT-H are employed to accelerate convergence and improve generalization.

Each variant of the HoverNet-based model is trained independently using identical training configurations. During inference, the output maps are aggregated and post-processed following the original HoVerNet protocol to compute instance-level masks and cell-type classifications.

Evaluation is carried out using Dice, Panoptic Quality (PQ), precision, recall, and F1-score metrics. We report results across different tissue types to assess the generalizability of each architecture.

Results

Baseline Performance

To establish a reliable benchmark, I trained the original HoVer-Net architecture on the PanNuke dataset using the standardized training pipeline described earlier. The baseline model achieved an NP-Dice score of 0.8686, indicating strong segmentation accuracy for nuclear regions. For type prediction (TP), Dice scores were 0.9703, 0.7978, 0.6728, and 0.6939 across the four nuclear subtypes, reflecting varying levels of classification performance.

In terms of instance-level evaluation, Panoptic Quality (PQ) scores were 0.4005 for neoplastic, 0.2830 for inflammatory, and 0.4526 for other cell types. Additionally, standard classification metrics such as F1-score, Precision, and Recall were computed to provide a comprehensive view of the model's strengths and weaknesses. These results serve as a foundational reference for comparing the performance of all proposed architectural variants in subsequent experiments.

CNN-Based Architectural Enhancements

Three key modifications were explored within the traditional CNN-based design framework: the incorporation of Squeeze-and-Excitation (SE) blocks, the integration of multi-head self-attention modules, and the use of an enhanced Dense-Block-based decoder. The most effective configuration, referred to as MSDHV-Net, combined all three enhancements.

Compared to the original HoVerNet, MSDHV-Net achieved consistent performance gains across multiple metrics. Specifically, the TP-Dice score for inflammatory nuclei increased from 0.6728 to 0.7162, while the TP-Dice score for the "other" cell category improved from 0.6939 to 0.7241. The NP-Dice and overall classification Dice scores also exhibited an average improvement of approximately 2% (See Table 1 and 2 for more details).

However, despite these enhancements, performance on neoplastic cells remained challenging. Both the PQ and F1 scores for this class experienced a slight decline of around 2% relative to the baseline, suggesting that additional strategies may be required to address the morphological variability and contextual ambiguity characteristic of neoplastic nuclei.

Transformer-Based Architectural Variants

The effectiveness of replacing the original CNN encoder with Transformer-based architectures was also evaluated. Three model variants were explored: HoverSwinNet, which directly employed the SwinViT encoder from the MONAI SwinUNETR implementation; HoVerIT, which integrated a customized Swin Transformer encoder into the original HoVerNet three-branch framework; and HoverViTNet, which substituted the encoder with a vanilla Vision Transformer (ViT).

All Transformer-based models underperformed compared to the CNN-based baseline. Among them, HoverViTNet demonstrated slightly better results than HoverSwinNet, but overall, the Trans-former variants exhibited a consistent degradation of 3–5% across Dice, Panoptic Quality (PQ), and classification-related metrics (See Table 1 and 2 for more details). These results suggest that while Transformer encoders offer strong theoretical advantages in capturing long-range dependencies, their practical integrations into multi-task nuclei segmentation frameworks remains a non-trivial challenge that warrants further investigation.

Summary of Comparative Performance

Among all proposed architectures, MSDHV-Net demonstrated the most consistent and mean-ingful improvements(See Figure 1 and 2 for comparision), particularly for inflammatory and other nuclear types. The transformer-based models, while conceptually promising, require further tuning or hybridization to outperform well-optimized CNN backbones on the PanNuke dataset.

Figures, Tables and Schemes

Table 1: Segmentation Metrics for Each Model (NP-Dice and TP-Dice)

Model	NP-Dice	TP-Dice-0	TP-Dice-1	TP-Dice-2	TP-Dice-3
HoverNet	0.8686	0.9703	0.7978	0.6728	0.6939
HoverNetEnhanced	0.8696	0.9697	0.8119	0.7175	0.7129
Multihead-HoverNet	0.8682	0.9698	0.7933	0.6643	0.6904
MSDHV-Net	0.8696	0.9697	0.8105	0.7162	0.7241
HoverSwinNet	0.8365	0.9643	0.7381	0.6221	0.6380
HoverViTNet	0.8564	0.9662	0.7672	0.6783	0.6816
HoverIT	0.8534	0.9676	0.7433	0.6297	0.6455
CellViT	0.8000	0.9729	0.9764	0.9632	0.9331

Table 2: Classification Metrics per Model (PQ, Recall, and Precision)

Model	PQ-1	Recall-1	Precision-1	PQ-2	Recall-2	Precision-2
HoverNet	0.4005	0.4192	0.4004	0.2830	0.2935	0.3206
HoverNetEnhanced	0.3843	0.3971	0.3924	0.2820	0.3119	0.2952
Multihead-HoverNet	0.3947	0.4070	0.4018	0.2813	0.2972	0.3150
MSDHV-Net	0.3825	0.3986	0.3872	0.2891	0.3179	0.3046
HoverSwinNet	0.3744	0.3824	0.3913	0.2400	0.2342	0.2945
HoverViTNet	0.3788	0.3984	0.3984	0.2793	0.3108	0.2976
HoverIT	0.3744	0.3984	0.3823	0.2400	0.3108	0.2976
CellViT	0.5606	0.6900	0.7200	0.4316	0.5700	0.5900

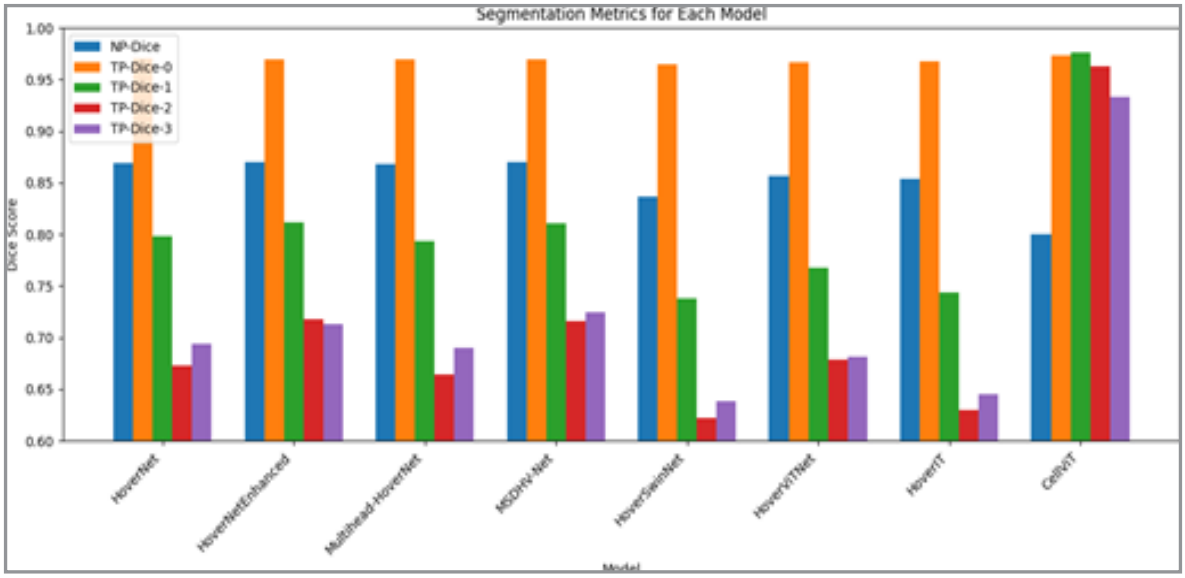


Figure 1: Bar Chart of Segmentation Metrics for each model.

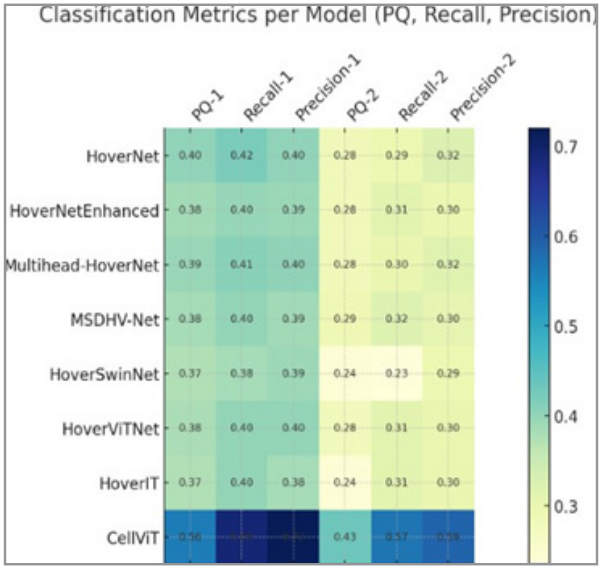


Figure 2: Heat map of Classification Metrics for each model.

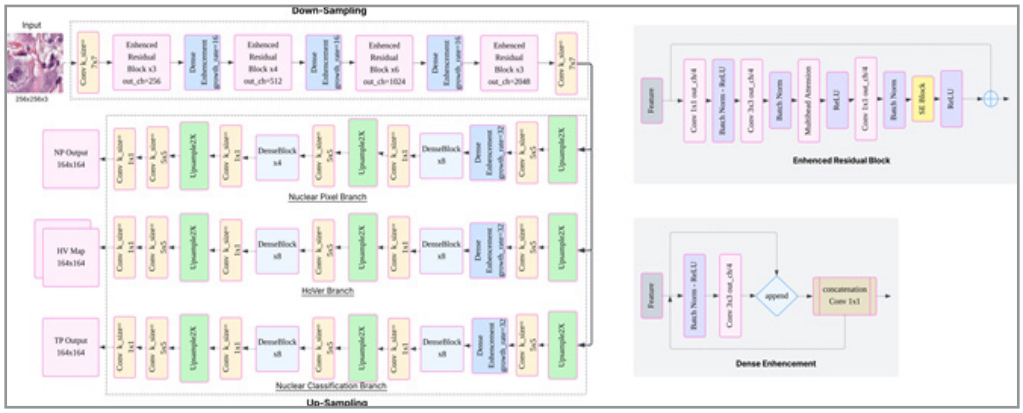


Figure 3: Overview of the MSDHV-Net architecture.

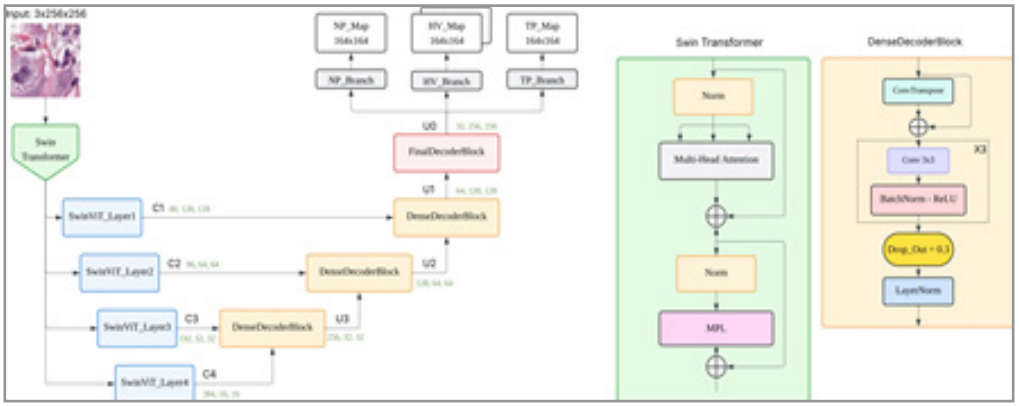
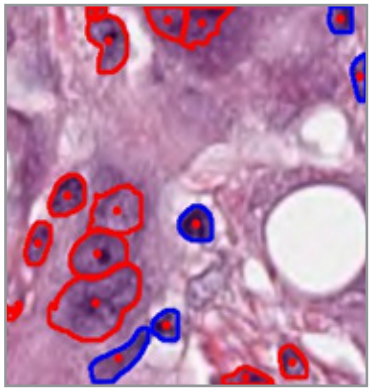
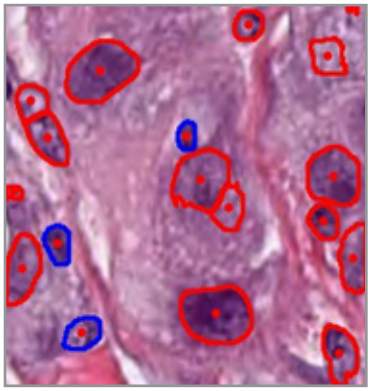


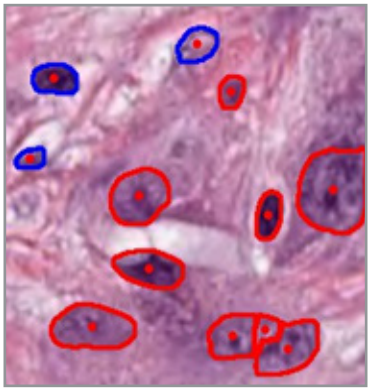
Figure 4: Overview of the HoVerIT architecture.



(a) HoverIT-1

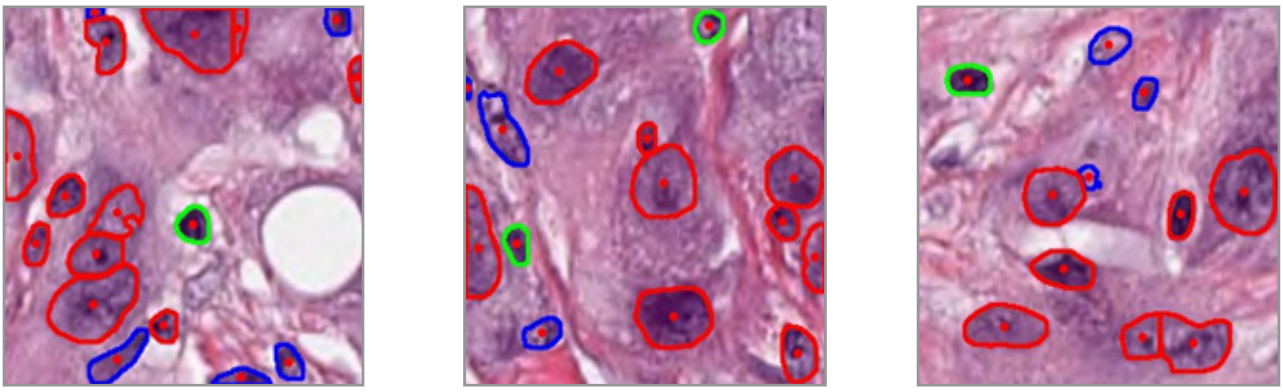


(b) HoverIT-2

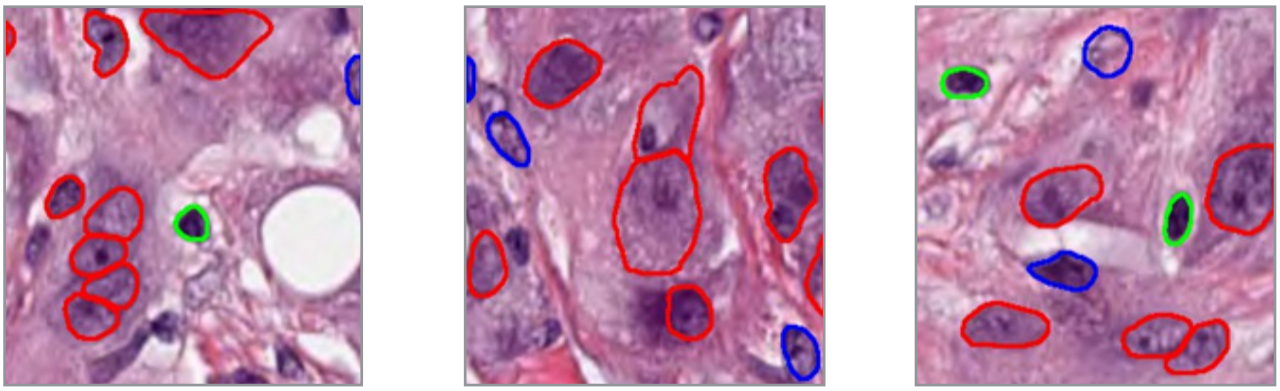


(c) HoverIT-3

(d) HoverSwinNet-1 (e) HoverSwinNet-2 (f) HoverSwinNet-3



(g) MSDHV-Net-1 (h) MSDHV-Net-2 (i) MSDHV-Net-3



(j) Ground Truth-1 (k) Ground Truth-2 (l) Ground Truth-3

Figure 5: Segmentation and classification results of three different models on multiple test patches.

Formatting of Mathematical Components
Dice Score:

$$\text{Dice} = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

X and Y represent the predicted and ground truth binary segmentation masks, respectively. |X| and |Y| denote the number of positive pixels in each mask, while |X ∩ Y| indicates the number of overlapping (true positive) pixels between them.

Binary Cross-Entropy (BCE):

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

N is the total number of samples. yi is the ground truth label (either 0 or 1) for the i-th sample, and pi is the predicted probability of the positive class for the same sample.

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2$$

N is the number of samples. yi and pi represent the ground truth and predicted values for the i-th sample, respectively. The loss penalizes the squared differences between predictions and true values.

Panoptic Quality (PQ):

$$\text{PQ} = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

TP, FP, and FN are the sets of true positive, false positive, and false negative instance predictions, respectively. IoU(p, g) denotes the intersection-over-union between a predicted instance p and its matched ground truth g.

F1 Score:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 Score is the harmonic mean of Precision and Recall, balancing the two metrics to provide a single measurement of accuracy.

Precision and Recall:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. Precision measures the proportion of correct positive predictions, while Recall measures the proportion of actual positives that were correctly identified.

Discussion

This study aimed to investigate architectural enhancements to the HoVer-Net framework for simultaneous nuclei segmentation and classification[5]. My findings demonstrate that thoughtful modifications within the CNN paradigm—specifically the integration of Squeeze-and-Excitation (SE) blocks, multi-head attention, and an enhanced DenseBlock decoder—can lead to measurable gains across multiple evaluation metrics. The best-performing model, MSDHV-Net (See Figure 3 for architecture), showed improved TP-Dice scores for Neoplastic, inflammatory and other nuclei types, indicating enhanced discriminative power in challenging classification scenarios.

These improvements are consistent with prior research emphasizing the importance of adap-tive channel recalibration (as in SE blocks) and attention mechanisms in deep feature extraction for biomedical imaging tasks. The enhanced decoder appears to better capture spatial dependencies and refine instance boundaries, especially in complex tissue environments.

In contrast, my exploration of Vision Transformer (ViT and SwinViT) encoders-although theoreti-cally appealing due to their global receptive field—did not yield superior performance. HoverSwinNet and HoVerIT (See Figure 4 for architecture) underperformed compared to both the baseline and CNN-enhanced models. One likely reason is that transformer-based models may require significantly larger training data or more domain-specific pretraining to outperform CNNs in medical imaging, for example the SAM-ViT pretrained model used in CellVit. Another limitation lies in integrating transformer encoders into multi-task frameworks like HoVer-Net, which may require carefully aligned intermediate feature representations.

Interestingly, all models struggled to improve classification metrics for neoplastic nuclei. This suggests either intrinsic ambiguity in their visual features or insufficient discriminatory signal in current feature representations. Future work could explore class-specific loss weighting or incorporate cell microenvironment context to enhance neoplastic classification.

Looking forward, promising directions include hybrid CNN–transformer architectures, domain- adaptive pretraining strategies, and exploring self-supervised representation learning to leverage unlabeled histopathology data. Additionally, integrating spatially aware attention mechanisms and refining decoder design may further enhance both instance segmentation and fine-grained classification capabilities.

Overall, this work highlights the value of selectively integrating modern deep learning compo-nents into established architectures, balancing innovation with task-specific constraints in computa-tional pathology.

Conclusions

In this work, I systematically investigated architectural enhancements to the HoVer-Net framework for nuclei instance segmentation and classification on the PanNuke dataset. Through the integration of SE blocks, multi-head attention, and a more expressive decoder, I developed MSDHV-Net, which consistently outperformed the original HoVer-Net in both segmentation and classification tasks—particularly for inflammatory and other nuclei types.

In contrast, transformer-based variants such as HoverSwinNet and HoVerIT did not demonstrate improved performance, suggesting that CNN-based backbones remain more robust and effective under the current dataset and training conditions. These results underscore the value of carefully engineered improvements to established CNN architectures, while also highlighting that successful

transformer integration—such as those attempted in CellViT with SAM-ViT pretraining-requires further domain adaptation and architectural refinement.

Overall, these findings provide practical insights for designing more accurate and efficient nuclei analysis models and lay a foundation for future work exploring hybrid CNN-transformer architectures and self-supervised learning approaches in computational pathology.

Author Contributions

Conceptualization, Shizhuo Qu; methodology, Shizhuo Qu; software, Shizhuo Qu; validation, Shizhuo Qu; formal analysis, Shizhuo Qu; investigation, Shizhuo Qu; resources, Shizhuo Qu; data curation, Shizhuo Qu; writing—original draft preparation, Shizhuo Qu; writing-review and editing, Shizhuo Qu; visualization, Shizhuo Qu; supervision, Shizhuo Qu; project administration, Shizhuo Qu. All authors have read and agreed to the published version of the manuscript.

Funding

This Research Received no External Funding.

Data Availability Statement

The dataset used in this study is publicly available. All experiments were conducted on the PanNuke dataset, which provides annotated histopathological images for nuclei instance segmentation and classification. The dataset can be accessed at https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke. No new

datasets were generated during the current study.

Acknowledgments

The author would like to express sincere gratitude to Jiang Yu from Yangtze River Delta Guozhi (Shanghai) Intelligent Medical Technology Co., Ltd., for entrusting the author with the responsibility of conducting the nuclei segmentation and classification component in the group project AI Multi-model System Prediction of Treatment Response to PD-1 Combined Chemotherapy in Advanced Gastric Cancer. This opportunity provided both inspiration and direction for the experimental work presented in this paper. Special thanks are also extended to Zhou Yin for technical support in server maintenance and PanNuke dataset preprocessing, and to Zhang Zihao (MSc, King's College London) for generously providing fundamental tutorials and guidance in computer vision for medical imaging. Their support has been instrumental in the successful completion of this study.

During the preparation of this manuscript, the author used ChatGPT (OpenAI, GPT-4, June 2025 version) to assist in refining academic language, correcting grammar errors, and improving writing clarity. The author has reviewed and edited all outputs and takes full responsibility for the content of this publication.

Conflicts of Interest

The author declares no conflicts of interest. This study was conducted during a cooperative research internship at Yangtze River Delta Guozhi (Shanghai) Intelligent Medical Technology Co., Ltd. While the company provided the research direction and access to computing resources, it had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sirinukunwattana, K., Snead, D., Epstein, D., Aftab, Z., Mujeeb, I., Tsang, Y. W., ... & Rajpoot, N. (2018). Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Scientific reports, 8(1), 13692.
2. Javed, S., Mahmood, A., Fraz, M. M., Koohbanani, N. A., Benes, K., Tsang, Y. W., ... & Rajpoot, N. (2020). Cellular community detection for tissue phenotyping in colorectal cancer histology images. Medical image analysis, 63, 101696.
3. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.
4. Isensee, F., Jäger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2019). Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128.
5. Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. (2019). Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis, 58, 101563.
6. Gamper, J., Koohbanani, N. A., Benes, K., Graham, S., Jahanifar, M., Khurram, S. A., ... & Rajpoot, N. (2020). Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778.
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
9. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., ... & Kleesiek, J. (2024). Cellvit: Vision transformers for precise cell segmentation and classification. Medical Image Analysis, 94, 103143.
10. Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., ... & Lu, C. (2023). Vision transformers for computational histopathology. IEEE Reviews in Biomedical Engineering, 17, 63-79.
11. Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. (2019). Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis, 58, 101563.
12. Zhang, Z. (2025). Enhancing Distributed Machine Learning through Data Shuffling: Techniques, Challenges, and Implications. In ITM Web of Conferences (Vol. 73, p. 03018). EDP Sciences.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
14. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. arXiv preprint arXiv:1905.09418.
15. Mao, X., Shen, C., & Yang, Y. B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Advances in neural information processing systems, 29.
16. Wang, Z., Li, T., Zheng, J. Q., & Huang, B. (2022, October). When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation. In European conference on computer vision (pp. 424-441). Cham: Springer Nature Switzerland.
17. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., & Gao, J. (2021). Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641.
18. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In International MICCAI brainlesion workshop (pp. 272-284). Cham: Springer International Publishing.
19. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
20. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
21. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
22. Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P.

(2019). Panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9404-9413).

Abbreviations

The following abbreviations are used in this manuscript:
MDPI: Multidisciplinary Digital Publishing Institute SE Squeeze-and-Excitation
CNN: Convolutional Neural Network
ViT: Vision Transformer

Appendix A. Additional Materials and Resources

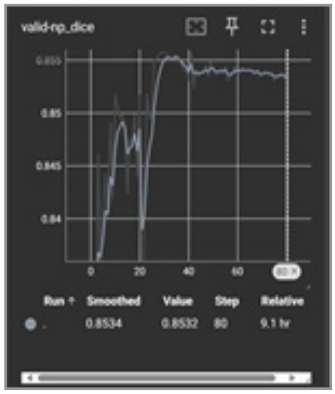
To promote reproducibility and transparency, additional materials are provided as follows:
• Trained model weights for all variants (e.g., HoverNetEnhenced, MSDHV-Net, HoverViTNet, HoVerIT) can be accessed at: https://drive.google.com/drive/folders/1fh0fiiGwIpPOafaoSF5 2WDrP4vAIxKY8?usp=drive_link.
• More overlay images of segmentation and classification results for representative samples are available at: https://drive.google.com/drive/folders/1urDlgA4QnI_25vAllJV2ouKhwg0Xdp5X?usp=sharing.

Appendix B. Model Structure and Implementation Code

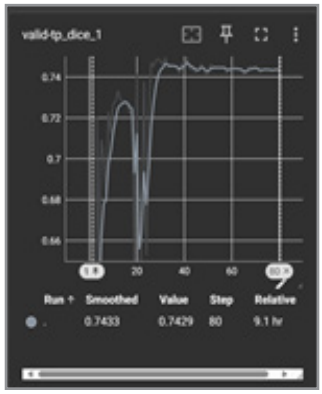
1. HoverSwinNet https://github.com/davidqu921/HoverSwinNet.
2. MSDHV-Net https://github.com/davidqu921/HoVer-Net-Enhenced.
3. HoVerIT https://github.com/davidqu921/HoVerIT.
4. HoverViTNet https://github.com/davidqu921/HoverViTNet.

These resources are intended solely for academic and non-commercial use.

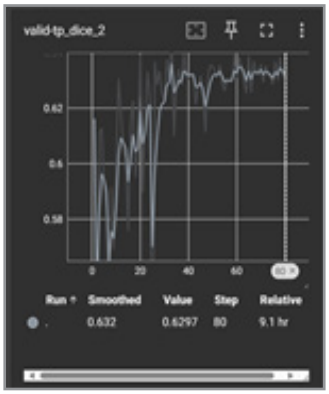
Appendix C: TensorBoard training logs and visualizations



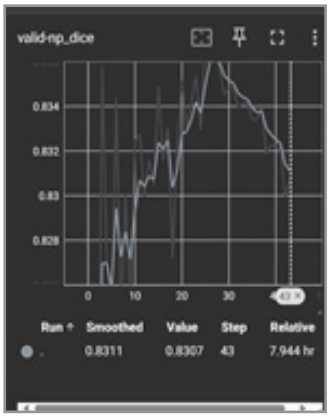
(a) HoverIT-valid-np



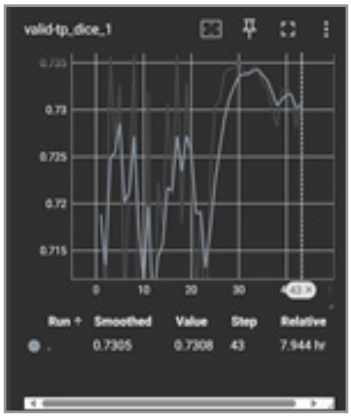
(b) HoverIT-valid-tp-1



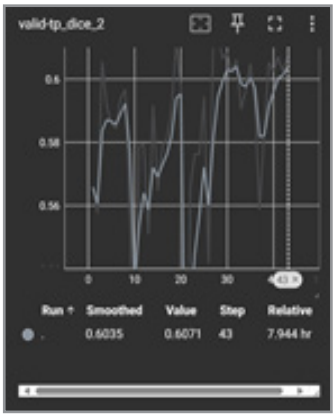
(c) HoverIT-valid-tp-2



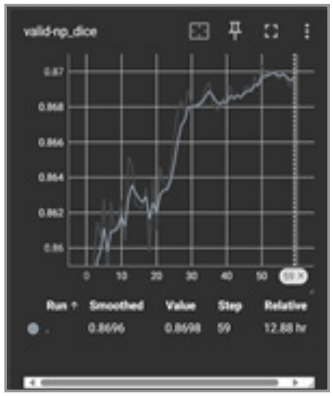
(d) HoverSwinNet-valid-np



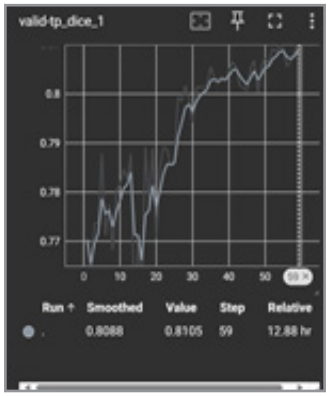
(e) HoverSwinNet-valid-



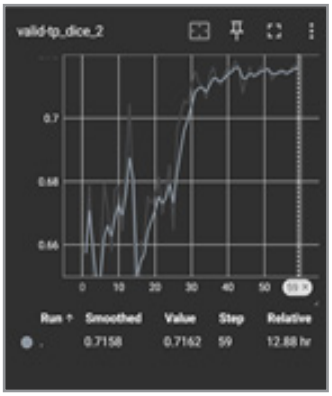
(f) HoverSwinNet-valid-tp- 2



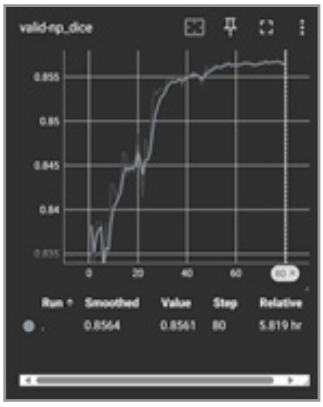
(g) MSDHV-Net-valid-np



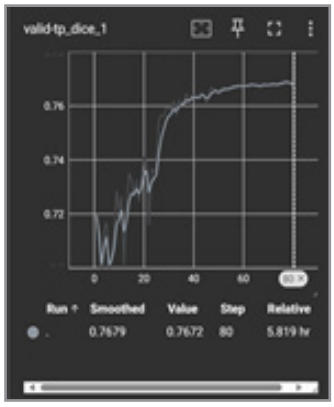
(h) MSDHV-Net-valid-tp-1



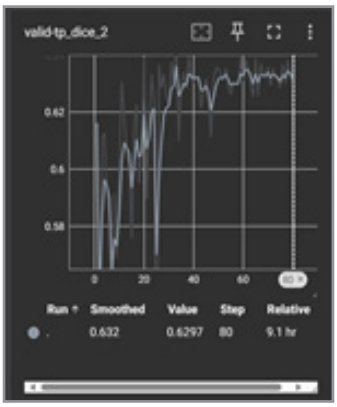
(i) MSDHV-Net-valid-tp-2



(j) HoverViTNet-valid-np



(k) HoverViTNet-valid-tp-1



(l) HoverViTNet-valid-tp-2

Figure A1: Segmentation and classification results of different models on multiple validation metrics