



Research on Frontiers

ARTICLE

Architectural Enhancements to HoVer-Net for Automated Nuclear Segmentation and Classification

Qu Shizhuo

DOI: <https://doi.org/10.62978/2601qsz120>



North American Chinese Research Press

Architectural Enhancements to HoVer-Net for Automated Nuclear Segmentation and Classification

Qu Shizhuo*

Department of Computer Science, University of Toronto Scarborough, Toronto, Ontario, Canada
davidsz.qu@mail.utoronto.ca

Abstract: Nuclear instance segmentation and classification support quantitative computational pathology by connecting cell boundaries with biologically meaningful labels. This exploratory study implemented five architectural variants of HoVer-Net on PanNuke, spanning channel recalibration, encoder attention, decoder refinement, and two transformer encoder integrations. The variants were evaluated using a shared reported image-patch partition and core metric pipeline, with one reported training run per configuration. The composite CNN configuration, MSDHV-Net, achieved inflammatory type-map Dice of 0.7162 versus 0.6728 for the reproduced HoVer-Net baseline, and merged Other type-map Dice of 0.7241 versus 0.6939. The SE-only variant yielded slightly higher neoplastic and inflammatory type-map Dice than the composite, while MSDHV-Net's neoplastic instance PQ was 0.3825 versus the baseline's 0.4005. In the reported setup, both transformer encoder substitutions had lower nuclear-pixel Dice than the reproduced baseline; HoverViTNet also showed a higher inflammatory type-map Dice than that baseline. Together, these point estimates describe how five implementation choices distribute performance across pixel overlap, cell type, and the available instance metrics. The study contributes an organized experimental comparison that identifies class-specific directions and concrete priorities for repeated, source-aware evaluation.

Keywords: Nuclear Instance Segmentation; Histopathology Image Analysis; Deep Learning; HoVer-Net

1. 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 microenvironment 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

Histopathology analysis requires more than a foreground mask. A semantic segmentation model can identify pixels that belong to nuclei, but it may merge adjacent objects and cannot necessarily assign a separate identity to each cell. Instance segmentation addresses this distinction by producing one region per nucleus. When type classification is added, the system must solve three interdependent problems: detect nuclear material, separate touching objects, and attach a categorical label to each recovered instance. An error at an early stage can propagate downstream; for example, an inaccurate boundary can alter both the measured morphology and the assigned type.

This dependency is important for quantitative tissue analysis. Counts, spatial distributions, and neighborhood relationships are computed from recovered instances rather than from isolated

*Corresponding Author

pixels.^{1,2} A model that achieves strong overall overlap may still be unreliable for a less frequent or morphologically variable category. For this reason, architectural evaluation should examine both region-level overlap and instance- or class-specific behavior. The present study adopts that multi-level perspective when comparing HoVer-Net⁵ variants.

Conventional image-processing techniques have limited robustness in these settings because handcrafted thresholds and shape assumptions are sensitive to staining variation, occlusion, and irregular morphology. Deep learning approaches, particularly convolutional neural network (CNN) models such as U-Net and nnU-Net, have substantially improved biomedical segmentation by learning hierarchical features directly from annotated data.^{3,4} HoVer-Net extends this development through a multitask design that simultaneously predicts a nuclear-pixel (NP) map, horizontal and vertical (HoVer) distance maps, and a nuclear-type (TP) map.⁵ The complementary outputs support separation of touching instances as well as assignment of type labels, making HoVer-Net a strong baseline for multi-tissue benchmarks such as PanNuke.⁶

Despite its effectiveness, the original HoVer-Net relies on a residual CNN encoder whose locality bias may limit the explicit modeling of long-range spatial relationships. Vision Transformer (ViT), Swin Transformer, and hybrid designs such as CellViT have demonstrated that self-attention can capture broad contextual dependencies relevant to heterogeneous histopathology images.^{7,8,9,10} However, replacing a CNN encoder within HoVer-Net is not a neutral substitution: transformer features must be aligned with three task-specific decoders, and their training behavior depends on data scale, pretraining, token resolution, and optimization. The value of transformer backbones in this particular multitask setting therefore requires controlled empirical evaluation rather than an assumption that a newer encoder will necessarily improve every output.

We therefore examine a portfolio of modifications at increasing levels of intervention. The first preserves the CNN backbone while testing SE channel recalibration, multi-head attention, and their combination with a denser decoder.^{11,13,15} The second integrates a vanilla ViT or a MONAI SwinUNETR encoder within the HoVer-Net output framework.^{7,18} Placing these alternatives alongside one reproduced HoVer-Net reference makes their distinct output profiles visible, even when no single configuration leads every measure.

The study is an exploratory implementation comparison on PanNuke, rather than a claim of clinical validation or a new attention mechanism. Its value lies in testing a deliberately varied set of established modules within HoVer-Net's three-output task, then examining how nuclear-pixel overlap, type-map overlap, and available class-specific instance metrics respond. The externally pretrained CellViT configuration provides context for the wider design space, not a matched architectural ablation.⁹

This study makes four connected experimental contributions:

(1) Channel recalibration. Squeeze-and-Excitation (SE) blocks were integrated into the CNN encoder to test whether adaptive emphasis of informative channels improves nuclear segmentation and type prediction.¹¹

(2) Attention-based encoders. Multi-head self-attention, a vanilla ViT encoder, and a MONAI SwinUNETR-based encoder were evaluated while preserving HoVer-Net's three-branch output logic.^{7,8,13,18}

(3) Composite decoder configuration. A deeper DenseBlock-based decoder with additional skip connections, dropout regularization, and convolutional refinement was implemented alongside SE and attention in MSDHV-Net.¹⁵ This tests the assembled configuration as a whole; a decoder-only comparison is reserved for a later isolated ablation.

(4) Cross-family implementation comparison. Five HoVer-Net variants were evaluated against the reproduced baseline using common core data and evaluation procedures. The comparison reports complementary outcomes for each design and distinguishes the separately initialized CellViT reference from these within-project variants.⁹

All HoVer-Net-based experiments used the same reported PanNuke partition, principal data pipeline, optimizer family, and evaluation code. The common setup supplies a practical reference for comparing implemented options; architecture-specific optimization and source-grouped validation remain separate questions. By presenting the full metric profile of every attempted modification, including configurations that lead different channels, the paper offers a concrete starting point for selecting the most informative follow-up experiments.

2. Materials and Methods

2.1 Dataset

All experiments used the publicly available PanNuke benchmark of annotated histopathology image patches spanning 19 tissue types.⁶ PanNuke originally distinguishes neoplastic, inflammatory, connective, dead, and epithelial nuclei. To focus the initial multi-output comparison on two clinically salient categories and one pooled residual category, we kept neoplastic and inflammatory nuclei separate and merged connective, dead, and epithelial nuclei as “Other.” This operational label is useful for the reported type-map experiment, although a pooled score cannot specify which of its three constituent classes produced a change.

The dataset was chosen for its comprehensive and high-quality annotations, as well as its established usage in prior research. These characteristics make PanNuke particularly suitable for evaluating multiclass nuclear segmentation methods and enable consistent comparisons with other published approaches.

PanNuke is particularly useful for this comparison because it combines tissue diversity with instance-level nuclear labels.⁶ A method must therefore accommodate substantial variation in staining, cellular density, object size, and nuclear appearance rather than optimize for a single organ. The remapped labels still define a multiclass task in which improvements may be uneven: features that help identify inflammatory nuclei may not improve neoplastic nuclei or the heterogeneous merged Other category to the same extent.

The dataset’s diversity also makes split consistency essential. If different variants are evaluated on different tissue mixtures, a change in metric value could reflect sample composition rather than model design. Accordingly, the comparison reuses a common partition and evaluation procedure. No new human-subject data or private clinical records were collected for this work; the experiments operate on the public benchmark and its existing annotations.

2.2 Preprocessing and Augmentation

To ensure a consistent comparison across all HoVer-Net-based models, we adopted the preprocessing and augmentation procedures from the original HoVer-Net 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

Before training, the nuclear-type labels were remapped to reduce the original five PanNuke categories to three biological categories. The original neoplastic and inflammatory labels were retained as classes 1 and 2, respectively, while connective, dead, and epithelial nuclei were

combined into class 3 (“Other”). Together with the background label (class 0), this produced four TP output channels for training and evaluation.

The study reports a fixed 6:2:2 partition of image patches for training, validation, and testing, reused for the HoVer-Net-based variants. The shared assignments keep the observed comparisons within one common patch composition. The available partition description does not establish grouping by source slide or patient, so the present test scores characterize the reported patch split. Recovering source identifiers and a split manifest would allow a subsequent source-aware evaluation and tissue- and class-level support reporting.

Preprocessing was held constant across the HoVer-Net variants. Images were normalized, resized when required by the implementation, and padded so that tensor dimensions remained compatible with encoder downsampling and decoder reconstruction. Labels underwent the corresponding spatial transformations to preserve pixel-to-instance alignment. The use of the same preprocessing code is important because apparently minor differences in normalization or padding can affect boundary quality and therefore alter Dice and PQ.

Training samples were shuffled to reduce ordering effects and improve exposure to heterogeneous tissues.¹² Any augmentation inherited from the project pipeline was applied consistently rather than tuned separately for a preferred variant. This choice favors comparability over model-specific optimization. It also means that an underperforming architecture may benefit from a different augmentation policy in future work, but such tuning falls outside the controlled comparison reported here.

2.3 Model Variants Based on the HoVer-Net Architecture

To examine a range of architectural modifications for nuclear instance segmentation and classification, five variants of the original HoVer-Net framework were implemented. Each retains the three principal outputs—`np_map`, `hv_map`, and `tp_map`—so that a common post-processing and metric pipeline can be applied.⁵ The variants span SE-based feature recalibration, encoder-level multi-head attention, a composite CNN configuration with an enhanced decoder, a ViT encoder, and a MONAI SwinUNETR encoder. This range provides two focused encoder changes, one assembled configuration, and two contrasting transformer substitutions. Available weights are listed in Appendix A; the verified implementation package belongs in Appendix B.

(1) HoVer-Net + SE (HoverNetEnhanced): This variant augments residual encoder units with SE blocks. Global channel descriptors are used to generate channel-wise weights that amplify informative responses and suppress less useful activations. The modification tests whether recalibration can improve representation without replacing the locality and translation bias of the convolutional backbone.¹¹

(2) HoVer-Net + Multi-Head Attention (Multihead-HoverNet): Multi-head self-attention (MHSA) modules are embedded in the encoder to model relationships between spatially separated features. Multiple attention heads can specialize in different interaction patterns, although their benefit depends on placement, feature resolution, and optimization.^{13,14} This variant isolates the effect of attention without simultaneously changing the decoder.

(3) HoVer-Net + SE + MHSA + Enhanced Decoder (MSDHV-Net): This composite CNN-based model includes channel recalibration and multi-head attention in the encoder and a deeper DenseBlock decoder with skip connections and convolutional refinement.^{11,13,15} It asks how the complete assembled configuration performs relative to the reproduced baseline and to the available single-module encoder variants. Because these components change together, its measured profile is attributed to the configuration, rather than individually to one module or a proven interaction.

(4) HoVer-Net + ViT Encoder (HoverViTNet): The residual CNN encoder is replaced with a ViT encoder that represents the image through patch tokens and models global relationships through self-attention.⁷ Transformer features are then routed into CNN-based decoding paths. This variant evaluates whether global token interaction compensates for the reduced convolutional inductive bias under the available data and training budget.¹⁶

(5) HoVer-Net + MONAI SwinUNETR Encoder (HoverSwinNet): This variant replaces the CNN encoder with the SwinUNETR backbone available through MONAI.¹⁸ Its multi-scale transformer features feed the three HoVer-Net decoders. Unlike the vanilla ViT substitution, this design supplies hierarchical features that can be connected to the NP, HoVer, and TP branches. It therefore probes a second practical encoder integration path within the same multitask output framework, without assuming that a common training budget optimizes both transformer configurations equally.

The variants span focused encoder changes, a composite CNN configuration, and two more extensive encoder replacements. HoverNetEnhanced and Multihead-HoverNet each alter one encoder mechanism; MSDHV-Net combines encoder and decoder changes; HoverViTNet and HoverSwinNet integrate distinct transformer feature hierarchies. This structured range permits a direct description of which attempted designs lead particular reported outputs. The two transformer rows also distinguish a vanilla patch-token approach from a hierarchical window-based integration. A decoder-only model and relevant pairwise combinations would be needed to determine the individual contribution of each MSDHV-Net component.

All variants retain the task decomposition that distinguishes HoVer-Net. The NP branch identifies nuclear pixels, the HoVer branch predicts directional distances used to separate touching instances, and the TP branch estimates nuclear categories.⁵ Maintaining these outputs permits use of a shared post-processing procedure. It also keeps the evaluation focused on representation and decoding rather than confounding architecture changes with a different definition of the task.

2.4 Training Configuration

The primary training setup used Adam¹⁹ with an initial learning rate of 0.0001 and ResNet-50 initialization²⁰ where architecturally applicable. Exploratory optimizer and learning-rate trials informed the selected setup, but were not a systematic architecture-specific tuning study. The common configuration gives the five variants a shared operational starting point. Transformer substitutions and the separately initialized CellViT reference have different optimization requirements; their observed scores are interpreted for the implementations and budget reported here.

An 80-epoch budget was specified, with batch size 16 on an NVIDIA L20 GPU with 44 GB VRAM. Appendix C displays selected validation-curve panels. Some visible panel endpoints differ from the stated budget; without the complete archived training histories, those endpoints cannot be assigned definitively to early stopping, checkpoint selection, or a truncated export. The reported scores are therefore described as outcomes of the available run records, with completed run lengths and checkpoint identifiers to be reconciled against those histories.

The loss functions followed the original HoVer-Net 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 (MSGF) for the horizontal-vertical distance (hv_map) regression task.

The shared budget makes the comparison operationally interpretable while leaving architecture-specific convergence as a follow-up question. In particular, CNN and transformer encoders differ in initialization and feature hierarchy, and an identical schedule need not place each

at its individually optimized point. The proposed comparison thus describes the tested implementations under the reported core settings. Confirming the complete training trajectories would further strengthen interpretation of differences among these variants.

The multitask loss reflects the distinct outputs. BCE and Dice terms encourage accurate nuclear-region and type-map predictions, whereas MSE and MSGE supervise continuous HoVer distance maps and their gradients. The gradient component is relevant to instance separation because abrupt directional changes help locate boundaries between touching nuclei. Loss definitions and weights were kept consistent across compatible variants so that observed differences were not intentionally induced by architecture-specific reweighting.

This study does not include a systematic benchmark of parameter count, throughput, memory consumption, or inference latency. Accordingly, the comparative conclusions concern predictive metrics; claims of unchanged or improved computational cost require a separate efficiency analysis.

2.5 Evaluation Metrics

Performance was evaluated using the Dice coefficient, Panoptic Quality (PQ), precision, and recall. Dice measures region overlap and is reported for the nuclear-pixel output and type-prediction channels. PQ combines instance-recognition and segmentation quality, making it sensitive to both missed or extra nuclei and the overlap of matched instances.²² Precision and recall provide complementary information about class-specific recognition errors. Metrics were examined by cell category wherever available because aggregate values can conceal class-specific gains and losses.

No single metric fully captures performance. NP-Dice rewards overlap of nuclear foreground but does not determine whether two touching nuclei were separated. TP-Dice evaluates overlap for type channels but can be influenced by class prevalence. PQ explicitly links recognition and segmentation by matching instances and penalizing false positives and false negatives.²² Precision and recall then show whether errors arise primarily from extra predictions or missed nuclei.

These complementary metrics motivate a profile-based reading of the tables. A higher TP-Dice may coexist with a lower PQ when improved type-map overlap does not carry through to post-processed instance matching. A recall increase accompanied by lower precision likewise indicates a different recognition balance. The available instance table covers neoplastic and inflammatory nuclei; the merged Other result is a type-map pixel-overlap result. Exact replication also calls for documentation of Dice aggregation, the PQ matching threshold, treatment of absent classes, and the unit used to aggregate per-image instance outcomes.

2.6 Definitions of Selected Metrics and Loss Terms

(1) 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 $|X \cap Y|$ indicates the number of overlapping (true positive) pixels between them.

(2) 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. y_i is the ground truth label (either 0 or 1) for the i -th sample, and p_i is the predicted probability of the positive class for the same sample.

(3) Mean Squared Error (MSE)

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

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

(4) Panoptic Quality (PQ)

$\text{PQ} = \Sigma \text{IoU}(p, g)$ over matched predicted and reference instances, divided by $(\text{TP} + 0.5 \text{FP} + 0.5 \text{FN})$.²²

TP denotes matched instance pairs, FP unmatched predictions, and FN unmatched reference instances; IoU is the overlap of each matched pair. The precise matching threshold, handling of absent classes, and cross-image aggregation used by the evaluation implementation should be recorded with the released evaluation configuration.

(5) Precision and Recall

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

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.

Together, these definitions separate pixel overlap, instance recognition, and classification error. Dice is used both as an evaluation measure and as part of the optimization objective, while BCE, MSE, and the gradient-based HoVer loss supervise different outputs during training. PQ, precision, and recall are evaluated after predictions have been converted into instances. This distinction matters because a lower training loss does not necessarily translate into a proportional gain in every post-processed metric.

2.7 Implementation and Reproducibility Controls

The shared data pipeline, sample assignments, label mapping, output branches, metric definitions, and post-processing code were maintained across the HoVer-Net-based variants. These controls make the reported profiles interpretable as within-project comparisons under common principal settings. The next level of reproducibility requires source provenance for the partition, exact configuration and environment files, and the evaluation command used to regenerate the table entries.

Each model contributed a single validation-selected checkpoint, which was then used for all metrics in that model's test row; metrics were not assembled from different checkpoints. Validation loss and PQ were monitored in the original process. The surviving manuscript description does not specify a single rule for combining these signals, and the selected epoch or step for every model must be confirmed from the archived logs. The tables consequently report single-run point estimates, which describe the observed rankings without estimating run-to-run uncertainty.

Appendix A links the available weights and illustrative overlays, while Appendix C displays selected training panels. A complete review package would join these with a verified split manifest,

seeds, software versions, preprocessing and augmentation settings, loss weights, selected checkpoint IDs, post-processing thresholds, and a single evaluation command. This consolidated record would allow the exploratory results here to be reproduced and extended as more targeted tests are added.

Multitask architectures require output channels and label indices to remain identical across training, inference, and evaluation. In this study, a fixed four-channel TP mapping was used throughout the pipeline: channel 0 represented background, channel 1 neoplastic nuclei, channel 2 inflammatory nuclei, and channel 3 the merged Other category comprising connective, dead, and epithelial nuclei.

3. Experiment

Experiments were conducted on PanNuke using the shared preprocessing and patch-based input pipeline described in Section 2. Every HoVer-Net-based variant was trained independently so that parameter updates and validation trajectories were specific to that architecture, while the principal data assignments and optimization settings remained fixed.

A CellViT configuration initialized with SAM-ViT-H weights was additionally considered as a contextual example of a different nuclei-segmentation pipeline.⁹ It uses a distinct architecture, pretraining source, and task-loss design from the five HoVer-Net variants. Accordingly, its score is discussed qualitatively and kept outside the principal controlled-result tables; comparisons intended to establish a performance ranking require matched partitions, labels, preprocessing, and evaluation code.

Each HoVer-Net-based variant was trained independently with the shared core configuration. At inference, its NP, HoVer, and TP maps were aggregated and post-processed following the project implementation of the HoVer-Net protocol to obtain instance masks and cell-type predictions. The reported test metrics for each model came from one saved validation-selected checkpoint. The exact selection rule and checkpoint identifier are to be reconciled with the archived training record, as described in Section 2.7.

Evaluation used Dice, PQ, precision, and recall. Reported results were inspected across type-prediction channels and available class-specific instance metrics. This evaluation strategy supports two distinct questions: whether a model delineates nuclear regions accurately and whether it assigns and separates instances reliably within individual categories.

All HoVer-Net-based variants used the same post-processing configuration for the predicted NP, HoVer, and TP maps. This consistency was necessary because thresholding, watershed-style separation, or instance matching choices can materially influence PQ and class-specific counts. The same output interpretation was used for the HoVer-Net variants so that comparisons reflect the learned maps rather than a customized post-processing advantage.

The experiment is best understood as a structured series of implementation trials rather than a complete factorial ablation. The SE-only and attention-only variants provide focused encoder comparisons; MSDHV-Net examines one assembled SE–attention–decoder design; and two transformer variants probe distinct feature hierarchies and integration paths. This range gives useful design-space coverage under one project pipeline. Isolating the revised decoder or interaction effects would require additional configurations and repeated runs.

4. Results

4.1 Baseline Performance

To establish a controlled reference, we 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 high nuclear-region overlap on the held-out test set. For type prediction (TP), the Dice scores were 0.9703 for background, 0.7978 for neoplastic nuclei, 0.6728 for inflammatory nuclei, and 0.6939 for the merged Other category.

At the instance level, the baseline produced a PQ of 0.4005 for the neoplastic category and 0.2830 for the inflammatory category in Table 2. Precision and recall were also computed because similar PQ values can arise from different error profiles. These baseline values establish a common reference for the architectural variants; they should not be interpreted as estimates of clinical performance outside the PanNuke test setting.

The baseline also illustrates why class-resolved evaluation is necessary. Its NP-Dice of 0.8686 is strong, yet TP-Dice varies from 0.6728 to 0.9703 across the four displayed TP channels. The spread indicates that reliable foreground segmentation does not imply uniform type classification. Categories with distinctive appearance or greater support may be easier to predict, whereas heterogeneous or underrepresented categories can remain difficult even when nuclear pixels are located correctly.

The four TP-Dice channels correspond to the remapped label space used in this study: TP-Dice-0 represents background, TP-Dice-1 represents neoplastic nuclei, TP-Dice-2 represents inflammatory nuclei, and TP-Dice-3 represents the merged “Other” category, which combines connective, dead, and epithelial nuclei. The class names are used consistently throughout the following analysis according to this mapping.

4.2 CNN-Based Architectural Enhancements

The CNN-based series covers SE channel recalibration, encoder-level multi-head attention, and their combination with enhanced DenseBlock decoding. HoverNetEnhanced changes SE alone, Multihead-HoverNet changes attention alone, and MSDHV-Net tests a composite SE–attention–decoder configuration. These rows permit comparison of their observed class-specific outcomes under the reported setup. They are not a complete factorial test of module contributions.

On the reported test partition, MSDHV-Net achieved NP-Dice 0.8696 compared with 0.8686 for reproduced HoVer-Net, a difference of 0.0010. Its type-map Dice was 0.7162 versus 0.6728 for inflammatory nuclei, and 0.7241 versus 0.6939 for the merged Other category. The larger type-map differences identify categories worth deeper investigation; the small foreground difference is simply a descriptive point estimate from one run. Other refers here to overlap in a pooled pixel-level type map, not to a demonstrated improvement in instance-level recognition for each constituent cell type.

The available instance metrics add an important dimension to this profile. MSDHV-Net's neoplastic PQ was 0.3825 versus the baseline's 0.4005, with recall 0.3986 versus 0.4192 and precision 0.3872 versus 0.4004. Its inflammatory PQ was 0.2891 versus 0.2830, with recall 0.3179 versus 0.2935 and precision 0.3046 versus 0.3206. The inflammatory result therefore describes a higher observed recovery of annotated cells alongside more unmatched predictions. The PQ difference of 0.0061 is not treated as a stable improvement without a variability estimate.

HoverNetEnhanced supplied another informative point in the CNN design space: NP-Dice 0.8696 and type-map Dice 0.8119 and 0.7175 for neoplastic and inflammatory nuclei, respectively. Both latter values are slightly above MSDHV-Net's corresponding scores. The attention-only

variant did not lead the displayed type-map channels. Together, these trials show why the composite should be evaluated alongside its available single-module encoder counterparts instead of presumed to inherit every benefit of its ingredients. Their differing class profiles help prioritize the next isolated tests.

MSDHV-Net remains a useful assembled configuration because it provides the highest reported merged Other type-map Dice among the controlled variants. The contrast with HoverNetEnhanced motivates a precise next question: does the revised decoder or attention placement alter type-map behavior differently across categories? The present data point to that question but cannot assign the pattern to a particular component. A decoder-only comparison, pairwise configurations, and repeated runs would supply the missing attribution.

4.3 Transformer-Based Architectural Variants

Two transformer implementations were explored: HoverSwinNet integrated the MONAI SwinUNETR encoder, whereas HoverViTNet used a vanilla ViT encoder with features routed into convolutional decoding paths.^{7,18} The pair offers a focused comparison of two concrete ways to supply transformer features to the three HoVer-Net outputs: hierarchical multiscale windows and global patch-token interactions. Their scores describe these integrations under the reported data handling and training records, rather than a general comparison of all CNN and transformer methods. This narrower design set also lets each substitution be read alongside its individual NP and type-map outcomes.

The observed NP-Dice values were 0.8365 for HoverSwinNet and 0.8564 for HoverViTNet, compared with 0.8686 for the reproduced CNN baseline. HoverViTNet reached inflammatory type-map Dice 0.6783 versus the baseline's 0.6728. Its foreground score and that single category score therefore moved in different directions, reinforcing the benefit of reporting multiple outputs. HoverSwinNet had lower values on the displayed type-map Dice channels under this setup. These observations identify which of the two implementations merits closer feature-scale and convergence analysis; they do not isolate the effect of tokenization, hierarchy, initialization, or optimization.

Several design considerations can guide that tuning. Transformer features may arrive at scales and representations that call for explicit adaptation to the existing HoVer and NP decoders. Decoder connections, token resolution, initialization, and task-specific pretraining are plausible variables for a subsequent matched study. The present trials identify the implementations for which those questions arise; they do not establish which proposed mechanism caused the observed scores.

CellViT illustrates a separate, purpose-built transformer pipeline with different initialization, losses, and architectural choices.⁹ It is retained as conceptual context, not a row in the within-project HoVer-Net tables or an estimate of what an encoder substitution alone would accomplish. A direct numerical comparison would need the same test identifiers, class mapping, preprocessing, and metric implementation. This distinction preserves the interpretability of the five implemented HoVer-Net alternatives.

4.4 Summary of Comparative Performance

The experiment yields a performance map rather than a single universal ranking. Among the HoVer-Net variants, MSDHV-Net had the highest merged Other type-map Dice (0.7241); HoverNetEnhanced had the highest neoplastic and inflammatory type-map Dice (0.8119 and 0.7175); and the baseline retained the highest displayed neoplastic PQ. Multihead-HoverNet did not lead those channels when attention was introduced alone. Figures 1 and 2 summarize the two

metric levels; Figures 3 and 4 connect the composite architecture with matched qualitative examples when the final images are inserted.

Matched qualitative overlays can complement the numerical profiles by showing where instances merge, split, or receive different type labels. Figure 4 is reserved for test patches displayed with a common scale and a baseline panel alongside the compared variants. The final illustration should identify tissue, contour colors, and how the examples were selected, including representative difficult cases. Such images aid interpretation; the reported test-set metrics remain the basis for the quantitative statements.

The NP branch and TP branch tell related but different stories. The 0.0010 NP-Dice difference for both SE-only and composite CNN variants shows closely spaced foreground scores in these individual runs. The more pronounced changes occur in selected type-map channels, while the available instance-level scores do not move uniformly with type-map overlap. This distinction is a substantive outcome of testing multiple configurations rather than compressing each model into one favorable average.

For a task emphasizing pixel-level inflammatory type overlap, the observed SE-only result slightly exceeds the composite (0.7175 versus 0.7162). For the pooled Other type map, MSDHV-Net has the largest reported score. For neoplastic instance PQ, the reproduced baseline has the higher displayed point estimate. These conditional comparisons tell a reader exactly which endpoint underlies each observation. They are most appropriately read as hypotheses for a repeated, source-aware analysis, not as guarantees for a new collection of slides.

All values in these tables come from one reported run per configuration. The emphasis is on direction, magnitude, and divergence across metrics under a shared project setup, particularly where a change in a type map does not imply better instance recognition. CellViT is discussed separately because its pretraining and task design prevent it from serving as an isolated HoVer-Net ablation. Reporting the numerical profiles in this way makes the exploratory evidence reusable for selecting focused follow-up configurations.

Table 1 Nuclear-Pixel and Nuclear-Type Dice for the HoVer-Net Variants

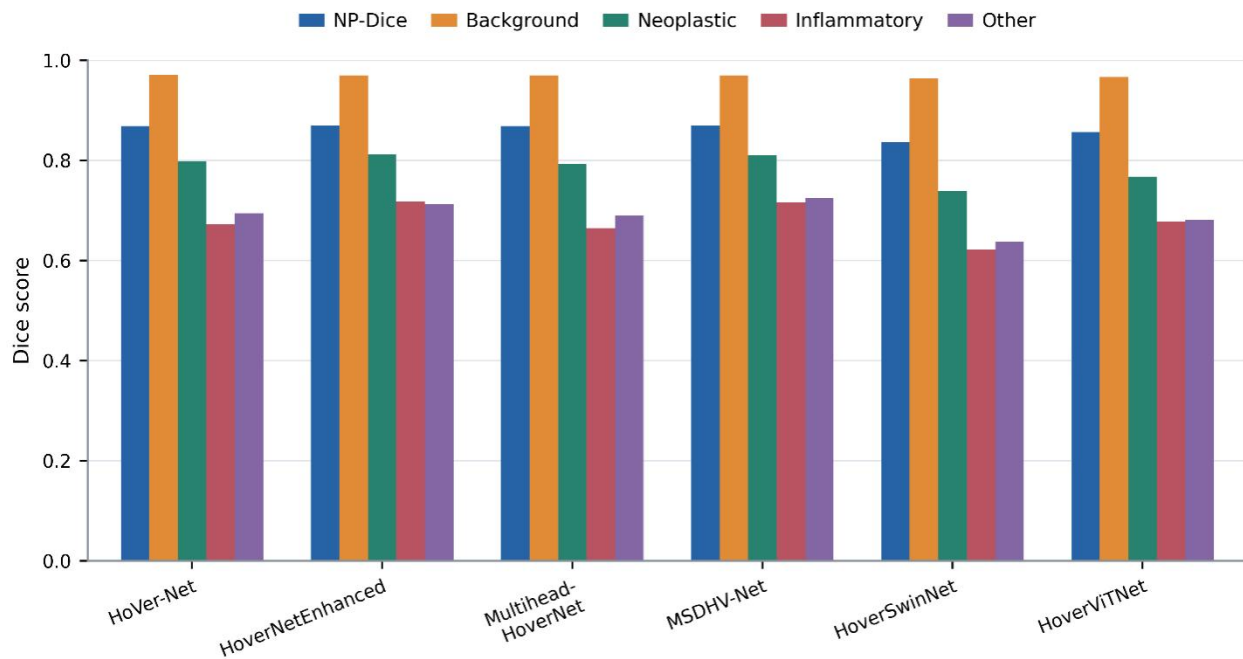
Model	NP-Dice	Background	Neoplastic	Inflam.	Other
HoVer-Net	0.8686	0.9703	0.7978	0.6728	0.6939
HoverNetEnhanced	0.8696	0.9697	0.8119	0.7175	0.7129
Multihead-HoverNet	0.8682	0.9698	0.7933	0.6643	0.6904
MSDHV-Net	0.8696	0.9697	0.8105	0.7162	0.7241
HoverSwinNet	0.8365	0.9643	0.7381	0.6221	0.6380
HoverViTNet	0.8564	0.9662	0.7672	0.6783	0.6816

Note: Type-map Dice channels correspond to background, neoplastic, inflammatory, and merged Other. Each value is a single-run point estimate on the reported test partition. CellViT is discussed separately because its architecture and initialization differ.

Table 2 Neoplastic and Inflammatory Instance Metrics for the HoVer-Net Variants

Model	Neoplastic PQ	Neoplastic recall	Neoplastic precision	Inflam. PQ	Inflam. recall	Inflam. precision
HoVer-Net	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

Note: PQ, recall, and precision are shown for neoplastic and inflammatory nuclei only; merged Other instance results were not available for this analysis. Each row is a single-run point estimate on the reported test partition.

**Figure 1 Nuclear-Pixel and Type-Map Dice Across the HoVer-Net Variants**

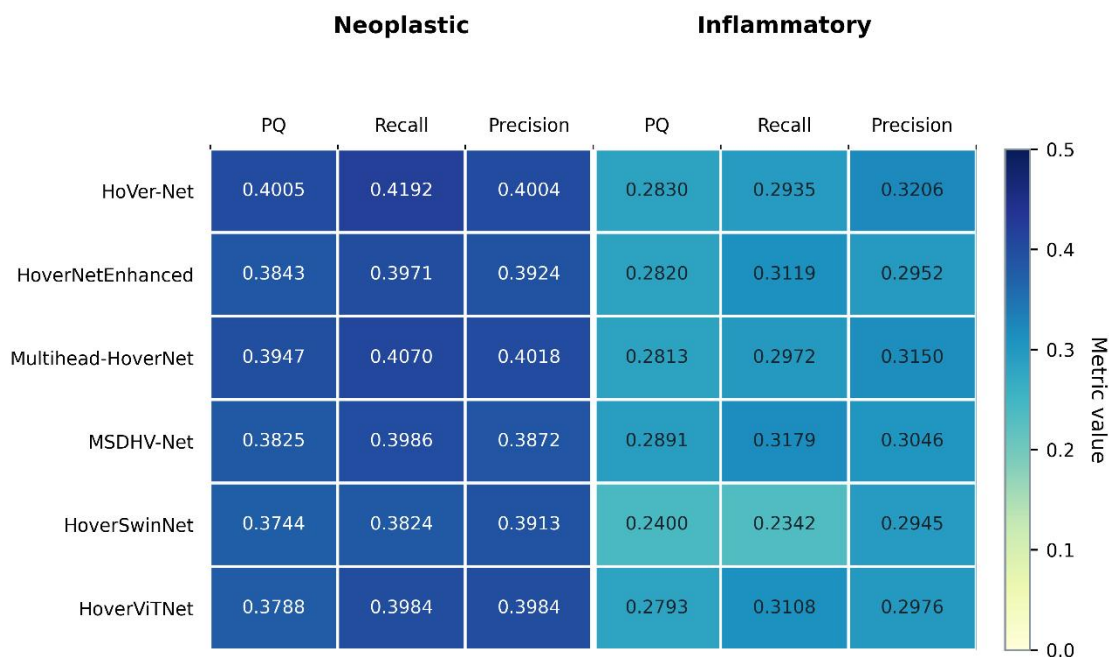


Figure 2 Neoplastic and Inflammatory Instance Metrics Across the HoVer-Net Variants

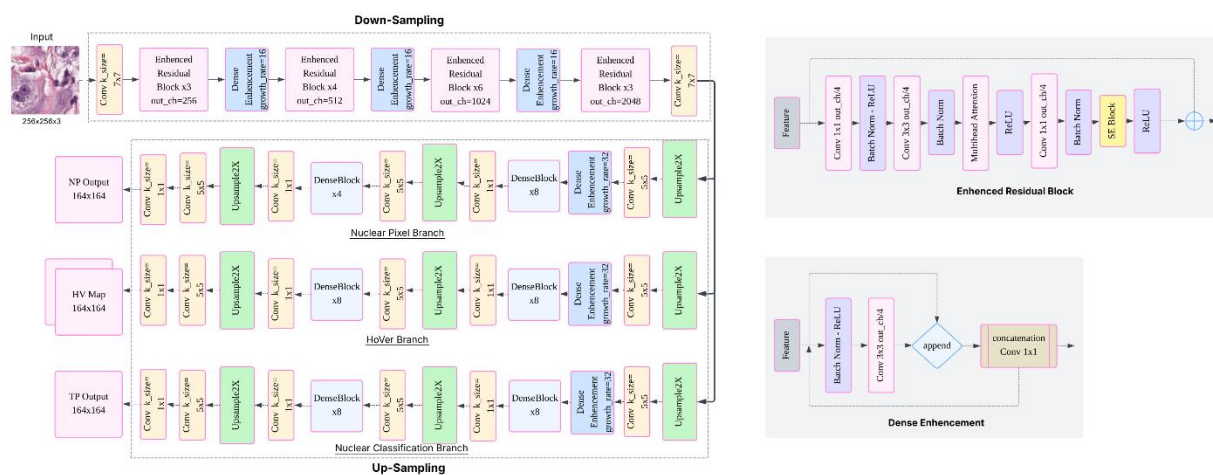
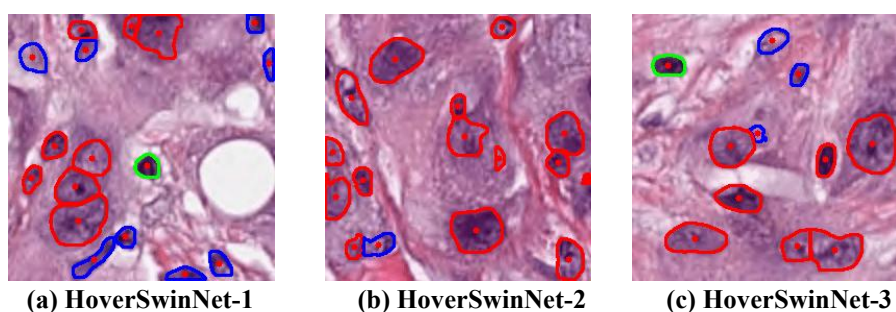


Figure 3 MSDHV-Net Composite Architecture



(a) HoverSwinNet-1

(b) HoverSwinNet-2

(c) HoverSwinNet-3

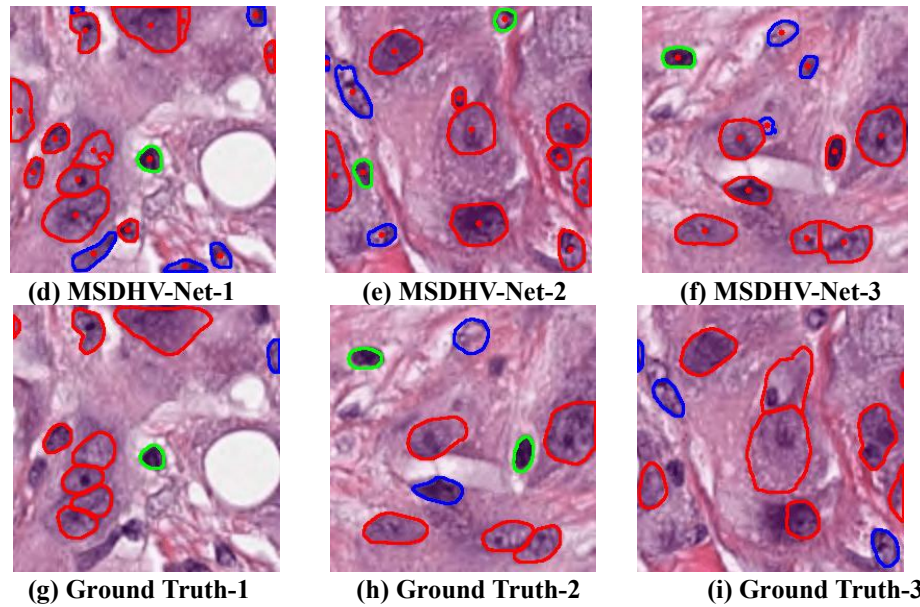


Figure 4 Matched Test-Patch Segmentation and Classification Overlays

5. Discussion

This exploratory study compared five ways of modifying HoVer-Net within a common project pipeline. The clearest finding is a differentiated metric profile: SE-containing CNN configurations reached higher type-map Dice for selected categories than the reproduced baseline in their individual runs, while neoplastic instance PQ and inflammatory precision did not follow the same pattern. This profile is useful because downstream objectives differ. Cell-type mapping, cell counting, and instance separation each favor careful attention to different endpoints, and a single aggregate performance label would obscure those choices.

The SE-only model is particularly informative within the trial set. Its neoplastic and inflammatory type-map Dice values were slightly higher than those of the composite MSDHV-Net, while the latter led the pooled Other type map. Channel recalibration offers a plausible representation-level explanation for some class-specific effects,¹¹ but this is a hypothesis to test through repeated and more isolated comparisons. The practical result is a clear pair of candidate configurations for a targeted next stage: SE-only as a focused encoder change and MSDHV-Net as an assembled system with a different category profile.

The revised decoder remains an interesting design element because preserving spatial detail can matter when adjacent nuclei must be separated. Dense connections and skip paths were introduced with that goal in mind. In MSDHV-Net, however, they were implemented together with SE and multi-head attention; the available experiments measure the assembled design rather than the decoder's independent effect. A decoder-only model and selected pairwise combinations would turn the present implementation survey into a cleaner component analysis. This distinction lets the current results stand as evidence about tested systems while defining a precise next experimental step.

The two transformer substitutions supplied a complementary set of results. Both had lower observed NP-Dice than the reproduced CNN baseline under the reported settings, while HoverViTNet had a slightly higher inflammatory type-map Dice. The result is specific to these feature connections, initializations, and training records. The contrast between a vanilla ViT and the SwinUNETR hierarchy motivates an integration-focused agenda: align intermediate scales

with the three decoders, examine branch-specific adaptation, and test matched pretraining and optimization. Existing purpose-built transformer systems such as CellViT occupy another part of the design space and should not be treated as the same intervention.⁹

The common stated epoch budget provided a practical comparison frame, and the selected validation panels offer an initial view of training dynamics. Complete logs are needed to establish which steps were completed and which checkpoint represented each model. That documentation would also permit a clearer comparison of convergence under architecture-appropriate schedules. The present scores remain informative descriptions of the saved runs, while a future tuned experiment could ask whether the ranking changes when each family receives a suitable optimization budget.

Neoplastic nuclei illustrate the value of preserving instance measures beside pixel measures. For MSDHV-Net, neoplastic PQ, precision, and recall were each below the reproduced baseline's corresponding score. That result points to a specific follow-up rather than erasing the type-map observation: inspect which nuclei were missed, split, merged, or assigned another category, then test whether class-sensitive learning and spatial context improve those failure modes. Morphology, tissue composition, and post-processing are plausible contributors that require analysis before one explanation is preferred.

Inflammatory nuclei show a different observed balance. MSDHV-Net recovered a larger fraction of annotated instances in the reported run, while its precision was lower, and its PQ difference from baseline was small. Some applications prioritize sensitivity to immune cells; others require reliable counts and therefore greater precision. Reporting all three metrics makes those priorities explicit and prevents the type-map Dice improvement from being read as a complete account of instance performance. The appropriate design can be selected only after the intended measurement target and acceptable error profile are specified.

The next stage can build directly on the configurations tested here. A focused decoder-only and pairwise ablation would probe the MSDHV-Net modules; repeated random seeds and paired test-set analysis at a verified independent unit would quantify uncertainty; original PanNuke labels and full Other-category instance metrics would clarify which cell types respond. A source-grouped and tissue-stratified partition, followed by an external cohort, would test whether the observed profiles extend beyond this reported patch split. These steps are sequenced research questions emerging from the multi-configuration survey, not requirements for interpreting the numbers as point estimates.

The scope of the present evidence is defined by one public benchmark, the reported fixed patch partition, and one available run per model. Related patches could share a source unless slide- or patient-level grouping is documented. Accordingly, the manuscript compares observed scores within this partition rather than estimating cross-source generalization or run-to-run stability. The useful contribution at this stage is the structured comparison of attempted implementations and the resulting class-specific measurement profile. Confirming source provenance and repeating key configurations would provide the next level of evidential support.

The four-channel type mapping was held fixed across the project, making the displayed variants comparable in this operational label space. Its pooled Other channel combines connective, dead, and epithelial nuclei, so the reported Other type-map result is not an outcome for those original classes separately. Table 2 also has no Other-category PQ, precision, or recall. The current analyses therefore discuss that channel specifically as pixel-level overlap. A class-resolved reanalysis could connect the observed pooled signal to biologically more specific instance outcomes.

The tables deliberately present the completed runs without confidence intervals. In particular, differences such as +0.0010 in NP-Dice or +0.0061 in inflammatory PQ are best used to generate targeted questions rather than strong ranking claims. Additional seeds would estimate training variation, while paired analysis on an independently grouped test set could estimate uncertainty over the evaluation units. Tissue-stratified summaries could then show whether a measured change recurs across the diverse PanNuke contexts or concentrates in a few tissues. The observed point estimates provide a starting map for designing those analyses efficiently.

An external evaluation would additionally ask how fixed checkpoints respond to differences in staining, scanners, preparation, and disease prevalence. Paired model overlays on the same test patches can make classification and boundary differences concrete, especially when both favorable and difficult examples are shown. Adding tissue identifiers, image scale, and a contour legend will strengthen the interpretation of Figure 4. These additions connect the numerical comparison to the actual image properties on which a follow-up investigation should focus.

The results position selective CNN modifications and direct transformer substitutions as a set of testable implementation choices, rather than as a universal hierarchy of architecture families. The SE-only and MSDHV-Net profiles invite a more precise study of channel recalibration, decoder design, and class-sensitive performance. The transformer profiles invite better feature alignment and matched optimization. By retaining the less favorable and mixed outcomes alongside the stronger type-map scores, the paper makes the design search more informative for readers building their own nuclei-analysis pipelines.

The public data, available weights, implementation descriptions, and selected training panels support inspection of the reported experiments. A consolidated anonymized package containing configuration files, split identifiers, complete logs, checkpoint IDs, and one evaluation command would make the comparison straightforward to rerun. These records would also resolve the visible discrepancies in training-panel endpoints and the exact checkpoint-selection rule. Appendix B identifies where the verified review-accessible implementation link belongs in the final submission.

A useful application-oriented reading follows from the endpoint differences. Immune-cell screening may emphasize inflammatory recall, whereas cell-density measurement depends on a different balance of precision and recall; neoplastic characterization may prioritize the relevant instance PQ and classification errors. The exploratory data cannot establish fitness for any clinical deployment, but it can identify which endpoints should be validated for a chosen task. This is more actionable than selecting a model solely by foreground NP-Dice.

More broadly, the collection of trials illustrates why architectural additions should be assessed in the full segmentation-and-classification pipeline. An encoder representation interacts with spatial reconstruction, three branch losses, label mapping, and post-processing. The study's main methodological contribution is to display these interactions through multiple complementary measurements of completed configurations, while keeping proposed mechanisms distinguishable from measured outcomes. It provides an empirical base on which isolated ablations and repeated, source-aware tests can be designed.

6. Conclusions

We implemented five architectural variants of HoVer-Net and compared their single-run outcomes on the reported PanNuke patch partition. SE-only and composite CNN modifications yielded higher type-map Dice in selected categories than the reproduced baseline. MSDHV-Net had the highest merged Other type-map Dice among the controlled variants, while HoverNetEnhanced had the highest neoplastic and inflammatory type-map Dice. The reproduced

baseline retained the higher neoplastic instance PQ. These conditional results form a comparative performance profile, not a single winner across the tasks measured.

The two tested transformer encoder substitutions had lower NP-Dice than the reproduced CNN baseline under the reported integration and training conditions; HoverViTNet nonetheless provided an informative type-map result for inflammatory nuclei. Together with the CNN series, these trials locate several concrete choices for the next iteration of the architecture: feature-scale adaptation, focused decoder tests, class-aware objectives, and matched initialization. MSDHV-Net is interpreted as one composite configuration because its three changes were implemented together.

The present study contributes a broad implementation survey within one task and evaluation pipeline. Repeated runs, source-verified partitioning, complete class-specific instance analysis, and full training metadata would establish the stability and provenance of the observed patterns. The available scores already specify where to direct those efforts: they distinguish foreground overlap from category overlap and instance recognition, and they identify both the promising directions and the trade-offs of the tested designs.

Architectural integration is therefore an experimental question with several observable answers. Under this reported setup, selective CNN refinements and the three direct transformer substitutions produced different, category-dependent performance profiles. The value of this initial comparison lies in making those profiles visible and in defining a tractable route from a varied set of implementation attempts to focused, reproducible follow-up experiments.

Abbreviations:

Table 3 Abbreviations Used in This Study

SE	Squeeze-and-Excitation
CNN	Convolutional Neural Network
ViT	Vision Transformer
NP	Nuclear Pixel
TP	Nuclear Type
PQ	Panoptic Quality

Data Availability Statement: The PanNuke dataset used in this study is publicly available at https://warwick.ac.uk/fac/cross_fac/tia/data/pannuke. No new dataset was generated during the study.

Acknowledgements: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Interest Statement: The author declares no conflict of interest.

References:

1. Sirinukunwattana, K., Snead, D., Epstein, D., Aftab, Z., Mujeeb, I., Tsang, Y. W., Cree, I., & Rajpoot, N. (2018). Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Scientific Reports*, 8, Article 13692.

2. Javed, S., Fraz, M. M., Epstein, D., Snead, D., & Rajpoot, N. M. (2018). Cellular community detection for tissue phenotyping in histology images. In *Computational Pathology and Ophthalmic Medical Image Analysis* (pp. 120–129). Springer.
3. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science* (pp. 234–241). Springer.
4. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18, 203–211.
5. Graham, S., Vu, Q. D., Raza, S. E., 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, Article 101563.
6. Gamper, J., Koohbanani, N. A., Graham, S., Jahanifar, M., Khurram, S. A., Azam, A., Hewitt, K., & Rajpoot, N. (2020). PanNuke dataset extension, insights and baselines. *arXiv*. <https://arxiv.org/abs/2003.10778>
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv*. <https://arxiv.org/abs/2010.11929>
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (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., et al. (2024). CellViT: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94, Article 103143.
10. Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., Madabhushi, A., & Lu, C. (2023). Vision Transformers for Computational Histopathology. *IEEE Reviews in Biomedical Engineering*, 17, 63-79.
11. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141).
12. Zhang, Z. (2025). Enhancing distributed machine learning through data shuffling: Techniques, challenges, and implications. *ITM Web of Conferences*, 73, Article 03018.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & 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*. <https://arxiv.org/abs/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, 2802–2810.
16. Wang, Z., Li, T., Zheng, J. Q., & Huang, B. (2022). When CNN meets ViT: Towards semi-supervised learning for multi-class medical image semantic segmentation. In *Proceedings of the European Conference on Computer Vision* (pp. 424–441). Springer.
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*. <https://arxiv.org/abs/2107.00641>

18. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop* (pp. 272–284). Springer.
19. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*. <https://arxiv.org/abs/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*. <https://arxiv.org/abs/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).

Appendix A

Additional Materials and Resources

Additional material accompanying the experimental comparison includes:

Trained model weights for all variants (e.g., HoverNetEnhanced, MSDHV-Net, HoverViTNet) can be accessed at:

https://drive.google.com/drive/folders/1fh0fiiGwIpPOafaoSF52WDrP4vAlxKY8?usp=drive_link.

More overlay images of segmentation and classification results for representative samples are available at:

https://drive.google.com/drive/folders/1urDlG4A4QnI_25vAlIJV2ouKhWg0Xdp5X?usp=sharing.

Appendix B

Model Structure and Implementation Code

Anonymized implementation and evaluation package: [insert verified review-accessible repository or supplementary link, including split identifiers, configurations, environment, checkpoints, and evaluation command].

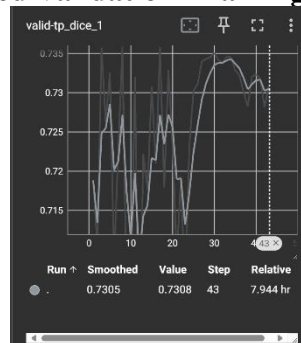
The available weights and implementation references provide material for research inspection; code reuse remains subject to the applicable repository licenses.

Appendix C

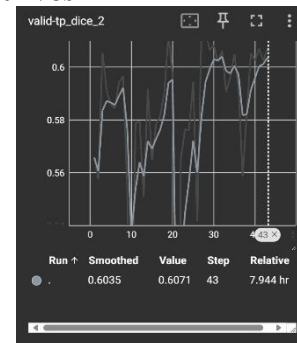
Selected Validation-Training Curves



(a) HoverSwinNet-valid-np



(b) HoverSwinNet-valid-tp-1



(c) HoverSwinNet-valid-tp-2

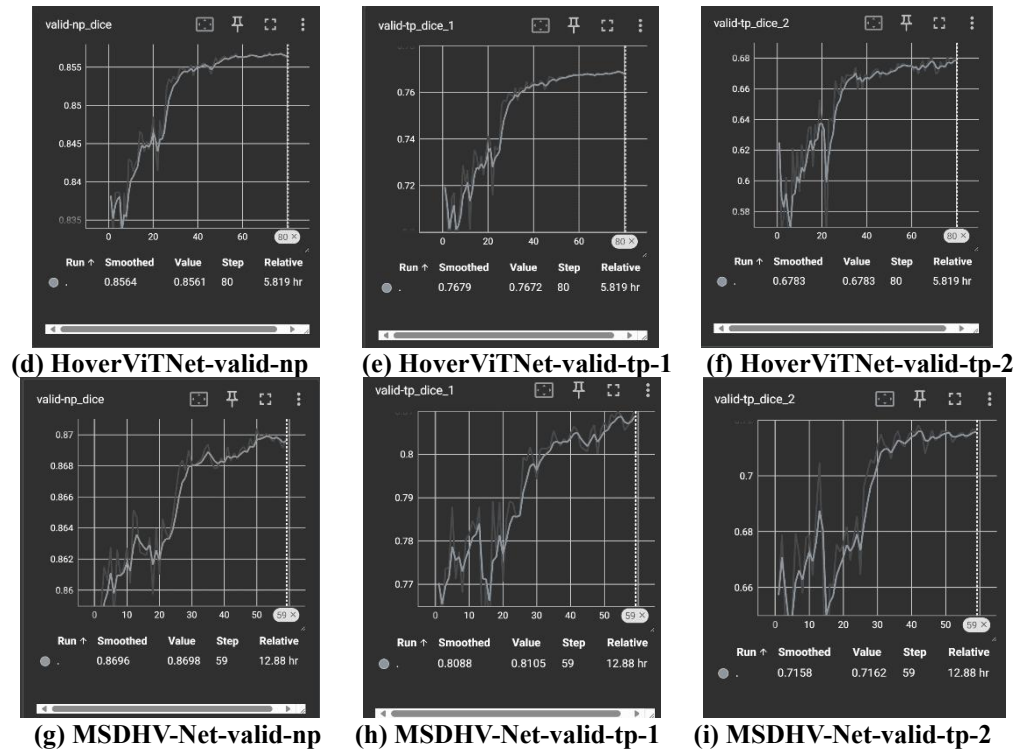


Fig. C.1 Selected Validation Curves for NP and TP Outputs of Three Model Variants. Complete run endpoints and selected checkpoint identifiers should be checked against the archived logs.