Computationally Efficient Stain Agnostic Nuclei Segmentation

Maria Kontaratou
Centralesupélec
Gif-sur-Yvette
maria.kontaratou@student-cs.fr

Zackarie Fredj - Zadoun
Centralesupélec
Gif-sur-Yvette
zacharie.fredj-zadoun@student-cs.fr

Supervisor: Stergios Christodoulidis
stergios.christodoulidis@centralesupelec.fr

Date: April 6, 2025

Abstract

Histopathology slides are essential for diagnosing many diseases, especially cancers. Their digitization into Whole Slide Images (WSIs) has enabled automated analysis, though their large size and staining variability present significant challenges. Hematoxylin and Eosin (H&E) staining is commonly used, while multiplexed imaging offers richer molecular insights but increases complexity.

Accurate segmentation of cells and nuclei in these images is crucial but complicated by staining differences, overlapping structures, and information loss when reducing input channels. To address this, we combined ChannelNet, which transforms multiplexed inputs into three-channel representations, with CellViT, a Vision Transformer-based model for nuclei instance segmentation.

Training was performed on the ZEISS and VECTRA subsets of the CPDMI 2023 dataset, with CODEX used as a held-out test set. Despite hardware constraints and a limited dataset, the model achieved good semantic segmentation performance. Instance separation remains challenging, but post-processing steps improved segmentation quality in dense or low-contrast areas. Code is available at: https://github.com/mkontaratou/stain-agnostic-nuclei-segmentation .

1. Introduction

Cancer poses a significant global burden, with millions of new cases annually, and is the second leading cause of death after cardiovascular diseases [1]. Despite advancements in non-invasive radiological imaging [2], microscopic tissue sample analysis remains a standard diagnostic procedure [3]. Pathologists identify tissue abnormalities to guide therapeutic approaches or further investigations. Analyzing cells and their distribution within tissues, such as detecting tumor-infiltrating lymphocytes or inflammatory cells in the tumor microenvironment, is crucial [4], but time-consuming and prone to high inter-observer variability.

The digitization of histopathology slides, such as hematoxylin and eosin (H&E) or multiplexed stained images, has enabled the application of computer vision (CV) techniques in cancer diagnostics. These technologies offer the potential for faster, more scalable, and more consistent diagnostic workflows, addressing the limitations of manual tissue analysis. Deep learning, particularly convolutional neural networks (CNNs), has been widely adopted in digital pathology tasks such as nuclei detection and segmentation, achieving clinical-grade performance in many settings [5][7]. More recently, transformer-based architectures - most notably Vision Transformers (ViTs) - have demonstrated superior performance across a range of image analysis benchmarks [8][9][10]. In the context of pathology, these models have begun to outperform traditional CNNs in tasks such as nuclei segmentation and tumor classification, establishing new standards for accuracy and generalization [9][10][11].

Despite these advances, analyzing multiplexed biomedical images presents unique challenges due to variability in staining, overlapping cell structures, and the complexity of integrating multiple biological markers. This work addresses these challenges by exploring two complementary approaches: ChannelNet [12], a channel-invariant segmentation framework, and CellViT [13], a ViT-based architecture for nuclei instance segmentation [10].

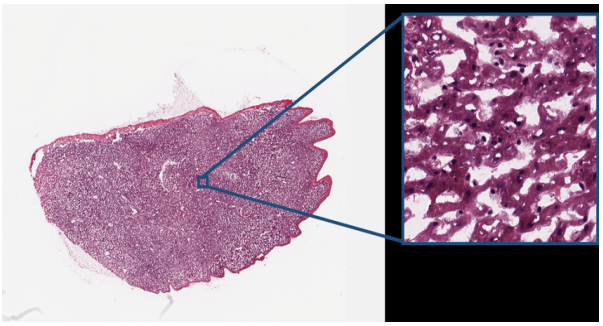

Figure 1. A digital pathology whole slide image. The 20,000 × 14,000 whole slide image is shown on the left at low magnification and a cropped region is shown on the right at high magnification

1.1. Definition

The nucleus is a critical organelle within eukaryotic cells, housing the cell's genetic material. It plays a pivotal role in cellular processes, including growth, division, and differentiation. Analyzing the nucleus can provide valuable insights into cellular health and function, making it a focal point in biomedical research.

Accurate identification and separation of nuclei in biomedical images are essential for quantitative analysis in digital pathology. Nuclei segmentation enables the extraction of morphological features, such as size, shape, and texture, which are crucial for diagnosing diseases and understanding cellular behavior. However, this task is challenging due to variations in nuclei appearance, overlapping boundaries, and the presence of artifacts in the images.

In multiplexed imaging, different biological markers or stains are captured in separate imaging channels. Each channel provides unique information about the spatial distribution and expression of specific biomarkers within the tissue. By integrating data from multiple channels, researchers can gain a comprehensive understanding of the cellular microenvironment. However, the complexity of managing and analyzing multiple channels poses significant computational challenges.

1.2. Research Problem

The primary research problem lies in developing robust and efficient methods for nuclei segmentation in multiplexed biomedical images. Current approaches often struggle with the variability and complexity of these images, leading to inaccuracies in segmentation. Additionally, the need to integrate information from multiple imaging channels adds another layer of difficulty. Addressing these challenges is crucial for advancing digital pathology and enabling more precise and reliable diagnostic tools.

By improving nuclei segmentation techniques and leveraging the rich information provided by multiplexed imaging channels, we aim to enhance the accuracy of cellular analysis and contribute to better diagnostic outcomes in biomedical research.

To address these challenges, we combined two approaches:

- **ChannelNet:** A channel-invariant deep learning architecture that generates a fixed three-channel representation of multiplexed images, regardless of the number or order of biomarkers. Integrated with the InstanSeg method, ChannelNet significantly improves the efficiency and accuracy of cell segmentation.

- **CellViT**: A deep learning architecture based on Vision Transformers for automated instance segmentation of cell nuclei in digitized tissue samples. CellViT eliminates the need for additional computational effort for feature extraction, achieving exceptional performance on the PanNuke dataset [6], which includes various tissue and nucleus types.

2. State of the art

2.1. ChannelNet: Enhancing Segmentation Performance by Optimizing Feature Utilization

Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated exceptional performance in biomedical image segmentation. However, a major challenge arises in handling multi-channel imaging data efficiently. In fluorescence microscopy and multiplexed imaging, individual biomarkers are captured in separate channels, leading to high-dimensional input data. Traditional CNN architectures often struggle with processing these multi-channel images due to increased computational costs and potential information loss when reducing the number of channels.

To address these issues, Gao et al. introduced ChannelNet, a method designed to optimize how multi-channel data is handled within segmentation models. The key motivation behind ChannelNet is to preserve the rich information contained in multi-channel inputs while maintaining computational efficiency. Instead of treating all channels equally, ChannelNet learns how to selectively merge and transform channels before feeding them into a downstream segmentation model. This approach ensures that the most relevant features are retained while minimizing redundancy, ultimately improving segmentation accuracy without significantly increasing computational complexity.

ChannelNet is designed as a learned channel mixer, meaning it automatically determines how to combine different imaging channels to maximize the segmentation model's performance. Instead of feeding raw multi-channel inputs directly into a segmentation model, ChannelNet first processes them using a lightweight neural network that learns an optimal feature representation.

The architecture of ChannelNet consists of three main components:

- **Channel Reduction and Mixing**: ChannelNet takes multi-channel images as input and projects them into a lower-dimensional representation while preserving critical information. This reduces the computational burden on the segmentation model.
- **Feature Transformation:** The method applies a set of learned transformations to enhance important features while suppressing irrelevant information. This helps the segmentation model focus on the most informative aspects of the input data.
- **Reconstruction and Integration**: The transformed features are then restructured into a format that can be efficiently processed by the segmentation network.

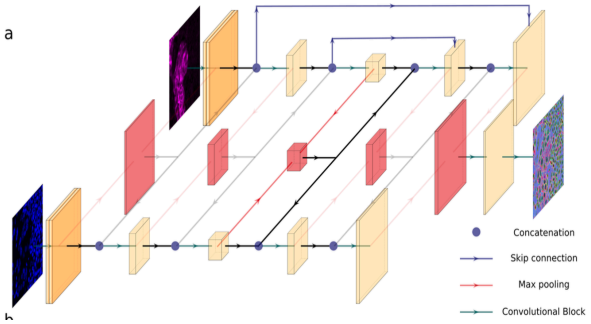

Figure 2. ChannelNet Architecture

Integrating ChannelNet into a segmentation pipeline significantly improves both accuracy and efficiency. Gao et al. demonstrated these benefits by evaluating the performance of InstanSeg, a state-of-the-art instance segmentation model, with and without ChannelNet. The results, obtained on the TissueNet dataset [15], highlight the impact of ChannelNet on segmentation performance.

Method	Target	F_1^{μ}	$F_1^{0.5}$	SQ	Time (s)	Images/second
Mesmer	Nuclei Cells	0.7115 0.6328	0.9030 0.8593	0.8421 0.8163	280	4.7
InstanSeg	Nuclei Cells	**0.7760** **0.6725**	0.9160 0.8699	**0.8738** **0.8343**	**31**	**42.7**
InstanSeg (+ ChannelNet)	Nuclei Cells	0.7649 0.6654	**0.9207** **0.8811**	0.8646 0.8252	36	36.8

Table 1. Quantitative segmentation results on the TissueNet test set

These results show that InstanSeg alone achieves strong performance, but when combined with ChannelNet, it exhibits improvements in $F_1^{0.5}$ (0.9207 vs. 0.9160 for nuclei, 0.8811 vs. 0.8699 for cells), indicating better segmentation accuracy.

A critical comparison is between InstanSeg with ChannelNet and a standard multi-channel approach without ChannelNet. On the CPDMI 2023 validation set, the F^{μ}_1 score for nuclei is 0.522 with ChannelNet versus only 0.498 when using a naive two-channel approach, and as low as 0.330 when using three channels without ChannelNet. This confirms that ChannelNet effectively preserves useful features from multi-channel inputs while reducing redundancy and noise, leading to better segmentation quality.

2.2. CellViT: Vision Transformers for Accurate Cell Segmentation and Classification

CellViT is a deep learning architecture based on Vision Transformers (ViT), designed for automatic nucleus segmentation in digitized tissue samples. Unlike traditional convolutional neural networks (CNNs), CellViT leverages the global attention mechanism of Transformers to capture complex spatial relationships in histopathological images. The CellViT architecture follows a U-shaped structure, combining a pre-trained ViT encoder-trained on a large-scale histological dataset (104 million image patches) with a symmetric decoder. This design allows CellViT to efficiently process high-resolution images, such as Whole Slide Images (WSI), extracting deep hierarchical features while simultaneously generating precise segmentation masks.

CellViT's performance has been evaluated on the PanNuke dataset, one of the most challenging benchmarks for nucleus segmentation, containing nearly 200,000 annotated nuclei across five clinically relevant classes and 19 tissue types. CellViT achieved a mean Panoptic Quality (PQ) of 0.50 and a detection F1 score of 0.83, outperforming previous models in both segmentation accuracy and nucleus classification. In comparison, HoVer-Net, a CNN-based model designed for simultaneous nucleus segmentation and classification, reported slightly lower results on the same dataset. Similarly, the U-Net architecture, while effective for biomedical image segmentation, did not achieve the same level of performance as CellViT on PanNuke. These results highlight the superiority of CellViT, particularly due to its use of Transformers and integration of foundational pre-trained models, for precise nucleus segmentation in complex histopathological images.

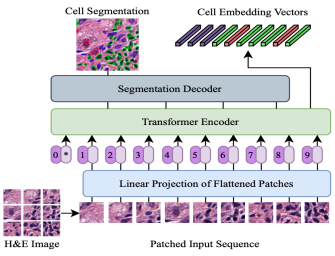

Fig.3 The CellViT Network

3. Our contribution

2.3. Dataset and Pre-processing

The Cross-Platform Dataset of Multiplex Fluorescent Cellular Object Image Annotations (CPDMI 2023 [14]) is designed for nuclear segmentation tasks in multiplex fluorescence microscopy images. This dataset includes acquisitions from three distinct imaging platforms – CODEX, ZEISS, and VECTRA – each exhibiting variations in signal intensity, contrast, and spectral characteristics.

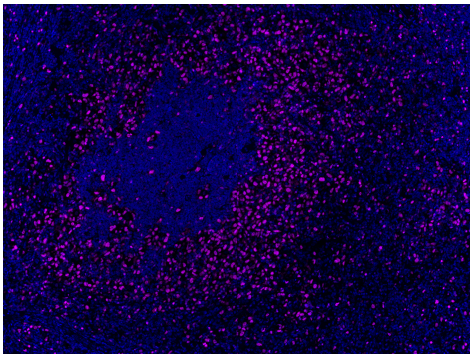

Figure 4. Full cell image from Vectra with DAPI coloration

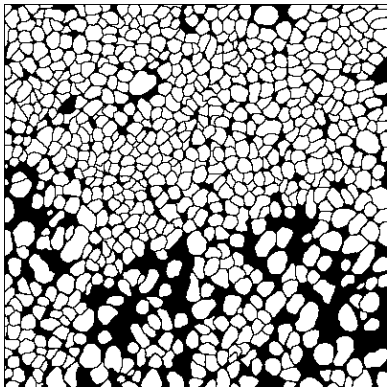

Figure 5. Crop mask of the Vectra image

For training data preparation, images from the ZEISS and VECTRA datasets were merged into a single dataset. Each sample consists of two types of images: a cytoplasmic/nuclear fluorescence image and a segmentation mask. To ensure a standardized input format for the model, we applied a uniform preprocessing pipeline. All images were converted to grayscale, resized to 256×256 pixels, and transformed into normalized tensors. A crucial aspect of our approach was the handling of channel numbers: since some images contained only a single channel, we adopted a channel replication strategy to obtain a 32-channel tensor. Specifically, for ZEISS images, the cytoplasmic image was duplicated to form a 2-channel tensor, which was then replicated 16 times to reach the required 32 channels. For VECTRA images, we combined the DAPI channel with the cytoplasmic image and applied the same replication strategy. The processed images and masks were then integrated into a PyTorch dataloader, enabling efficient loading in mini batches of 3 samples.

2.4. Channel Adapter Bridging 32-channel inputs and 3-channel models

A key design element of our implementation is the Channel Adapter module, introduced to reconcile the mismatch between the 32-channel input tensors used in our datasets and the expected 3-channel input of standard backbone architectures. The adapter comprises two 3×3 convolutional layers, each followed by batch normalization and ReLU activation, concluding with a 1×1 projection layer to reduce the dimensionality from 32 to 3 channels. A residual connection is conditionally applied when the input and output channels match, helping preserve spatial detail and improve gradient flow. This structure enables the model to learn a semantically meaningful projection from the repeated low-dimensional channels, rather than relying on naive duplication or hardcoded mappings. Furthermore, the spatial context captured by the 3×3 filters support the detection of fine-grained textures and edge features critical for nuclear segmentation. Despite incorporating batch normalization and residual connections for training stability, the Channel Adapter remains computationally efficient due to its shallow depth, low parameter count, and early placement in the network. By inserting this lightweight, trainable module at the input stage, we preserve compatibility with pretrained models while enhancing generalization on high-dimensional biomedical inputs.

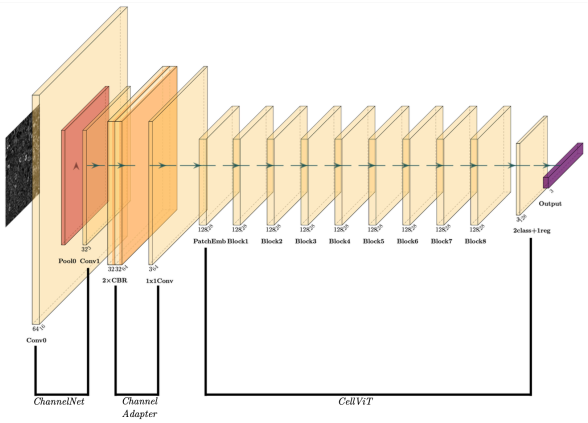

Figure 6. The ChannelNet_CellViT Architecture

2.5. Training

To ensure robust evaluation, training was exclusively performed on the VECTRA and ZEISS subsets of the CPDMI 2023 dataset, totaling 65 multiplexed samples, while CODEX was held out as an independent test set to assess generalization across staining protocols and acquisition domains.

The training pipeline incorporated a data preprocessing stage where input images were resized to 256×256, intensity-normalized using a Percentile Scaling transformation, and expanded to 32-channel tensors to match the input requirements of ChannelNet. Augmentations such as rotation and flipping were applied to enhance model robustness and reduce overfitting on small sample sizes.

Model optimization was carried out using the Adam optimizer with an initial learning rate of 0.001, adaptively reduced by 50% upon a plateau in validation performance (patience = 5 epochs). Due to hardware limitations, we adopted a batch size of 3 and implemented gradient accumulation (4 steps) to stabilize gradient updates and simulate a larger batch size. Training was accelerated with Automatic Mixed Precision (AMP) to improve computational efficiency.

Throughout our experiments, we evaluated a range of loss functions to improve segmentation performance. Dice Loss, though commonly used in biomedical image analysis, proved unstable - particularly in regions with clustered or overlapping nuclei - and led to suboptimal convergence. Tversky Loss, while designed to address class imbalance, did not yield substantial improvements without intensive hyperparameter tuning. Focal Loss, which emphasizes harder examples, tended to over-penalize during training, impairing generalization. Ultimately, the best performance was achieved using a combination of Lovász-Hinge and Cross-Entropy losses. The Lovász-Hinge [16] component is particularly advantageous in segmentation tasks with class imbalance, as it directly optimizes the Intersection-over-Union metric in a tractable form.

Combined Loss: $\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{Lovász}} + (1 - \lambda) \cdot \mathcal{L}_{\text{CE}}$

where

$\lambda = 0.7$,

$\mathcal{L}_{\text{Lovász}} =$ Lovász-Hinge loss on foreground,

$\mathcal{L}_{\text{CE}} =$ Cross-Entropy loss

The training dynamics observed in our setup follow a predictable yet informative pattern for nuclei segmentation. During the first 30–40 epochs, the model rapidly learns to distinguish nuclei from background, as reflected by a sharp decrease in both training and validation loss, and a steep increase in F1_mean. This indicates early success in segmenting well-defined nuclear structures. However, from epoch 40 onward, the F1_mean plateaus, suggesting that the model has saturated on easy examples and is now limited by harder cases such as: overlapping nuclei, low-contrast boundaries, and cross-domain variations from DAPI, Cell, and CODEX images. Interestingly, while F1_mean stabilizes, the F1@0.5 metric reaches almost 1.0 and remains steady, demonstrating that the model can correctly identify the presence of most nuclei, but struggles to finely delineate individual instances across tighter IoU thresholds.

This behavior is further validated by the Segmentation Quality (SQ) score, which stabilizes around 0.63: a respectable result for dense, complex histological images. The model also demonstrates computational efficiency: full training on 130 epochs completed in under 20 minutes, thanks to gradient accumulation and mixed-precision optimization. Overall, these learning curves and metrics suggest that the model is well-calibrated and converges effectively, with room for improvement in fine-grained instance separation, to be addressed in future work

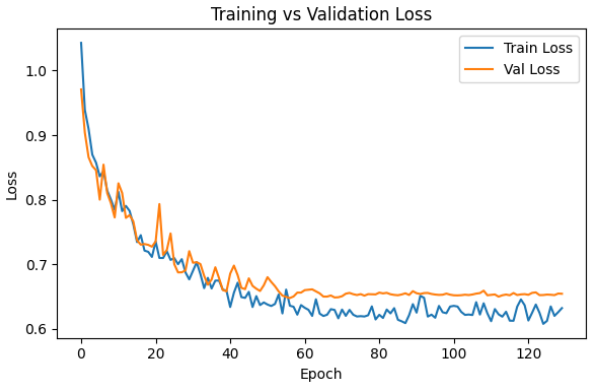

Figure 7. Comparison of training and validation performances of our mode.

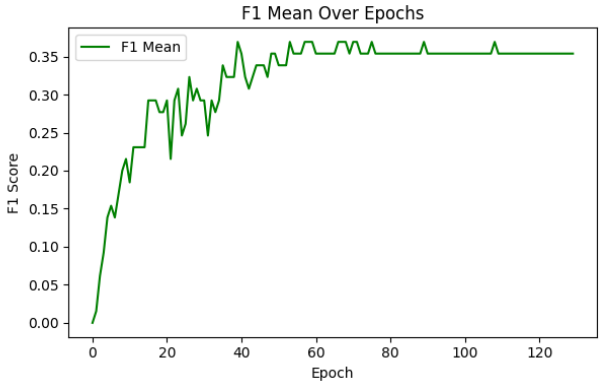

Figure 8. F1 Mean over epochs

Although the model predicted dense binary segmentation maps, accurate instance-level delineation required refinement to resolve merging of adjacent nuclei, suppress noise, and enhance structural coherence.

To address this, we implemented a morphological post-processing pipeline, consisting of:

- Low-threshold activation: Instead of using a hard 0.5 threshold, we applied a 0.3 threshold on the foreground probability map, derived from the softmax output. This decision was empirically validated to improve recall by preserving faint or partially stained nuclei that would otherwise be missed.
- Morphological filtering: We employed remove_small_objects and remove_small_holes operations to eliminate small noise regions (min size = 32 pixels) and close internal holes within segmented nuclei, enhancing the completeness and quality of the masks.
- Connected component labeling: Finally, we applied connected component analysis to assign unique instance IDs to each contiguous region. This was essential for computing instance-aware metrics such as Panoptic Quality (PQ) and Segmentation Quality (SQ), which penalize merged or split nuclei.

This post-processing stage significantly improved the coherence of the segmentation masks and directly contributed to more reliable evaluation results, particularly in low-contrast or high-density regions where the model predictions alone were insufficient for instance-level separation.

All experiments were executed on nodes sh[11–19] of our GPU cluster, each with an 8-core Intel Xeon® W (2125/2225) CPU, 32 GB RAM, and either a GeForce 1080Ti (11 GB) or GeForce 3080 (10 GB) GPU. Due to the high memory demands of our 256×256, 32-channel inputs, a batch size of 16 was infeasible. We therefore used a batch size of 3 with 4-step gradient accumulation and AMP for efficiency.

2.6. Validation

Method	$F_{\mu 1}$	SQ
Your Model (CellViT + ChannelNet)	0.354	0.633
InstanSeg (+ ChannelNet)	0.522	0.767
InstanSeg - two channel	0.498	0.761
InstanSeg - three channel	0.330	0.707
InstanSeg (+ ablated ChannelNet)	0.491	0.757

Table 2. Comparison of our model against baselines on the CPDMI 2023 Validation set

This comparison table highlights how the hybrid ChannelNet + CellViT model performs against several baselines on the CPDMI 2023 validation set, using the $F_{\mu 1}$ Score and Segmentation Quality (SQ) as evaluation metrics.

While our model achieves a respectable SQ of 0.633, indicating a solid ability to segment and reconstruct nuclear structures, its F_1^{μ} Score (0.354) lags other methods, particularly the ChannelNet + Instanseg model, which reaches 0.522 for F_1^{μ} and 0.767 for SQ. This suggests that although our method can accurately delineate segmentation masks, it struggles more with distinguishing individual nuclear instances, especially in densely packed or noisy regions.

Interestingly, models trained with two or three channels also outperform ours in F_1^{μ} while maintaining high SQ, indicating that more targeted input configurations or stronger instance separation mechanisms (as in InstanSeg) may be more effective for this task. Additionally, the ablation study on ChannelNet in InstanSeg shows relatively minimal performance degradation, highlighting its robust backbone.

2.7. Post-Processing

To further enhance segmentation accuracy, we implemented a lightweight morphological post-processing step aimed at refining instance-level predictions:

- **Softmax activation and thresholding:** The probability scores produced by the model were thresholded at 0.3, a value optimized to increase sensitivity for faint nuclei while maintaining precision. This step ensured that weakly fluorescent nuclei were not discarded.

- **Morphological cleanup:** Small, isolated detections caused by imaging noise and artifacts were removed to prevent false positives. Additionally, small gaps inside segmented nuclei were filled to improve mask coherence and ensure complete object detection.

- **Connected component labeling:** To enable accurate instance segmentation, individual nuclei were assigned unique labels, ensuring that touching or overlapping nuclei were correctly identified as distinct objects. This step was particularly important for evaluating Panoptic Quality (PQ) and Segmentation Quality (SQ), as it prevented the merging of closely packed nuclei into a single entity.

These post-processing refinements significantly improved the final segmentation outputs, particularly in addressing fragmented and incomplete nuclei.

3. Test & Results

Final model evaluation was conducted on the CODEX test set, comprising 9 previously unseen multiplexed microscopy images. Inference was performed with a batch size of 3 and a calibrated probability threshold of 0.3 to boost recall for faint nuclear signals. Post-processing steps (including artifact removal and instance separation) were applied to enhance segmentation quality. Performance was assessed using a comprehensive set of metrics: Pixel Accuracy, F1 Score, Precision, Recall, Mean Average Precision (mAP), as well as instance-level indicators like $F_1^{0.5}$, $F\mu_1$, Segmentation Quality (SQ), and Panoptic Quality (PQ).

Metric	Value
Panoptic Quality (PQ)	0.0363
Segmentation Quality (SQ)	0.5897
$F_{0.51}$ Score	0.0342
$F_{\mu 1}$ Score	0.0162
F1 Score	0.7147
Precision	0.8179
Recall	0.6368
Mean Average Precision (mAP)	0.7712
Pixel Accuracy	0.6548

Table 3. Test set evaluation metrics of our model

The model achieved high precision (81.79%), reflecting a low false positive rate, and a global F1 Score of 0.7147, suggesting a strong balance between sensitivity and specificity. A mAP of 0.7712 further supported its robustness across thresholds. However, instance segmentation proved more challenging (evidenced by PQ = 0.0363 and $F_1^{0.5}$ = 0.0342) highlighting the difficulty of delineating overlapping or tightly clustered nuclei. These outcomes indicate the model's strength in semantic segmentation, while also pointing to future improvements needed for instance-level accuracy.

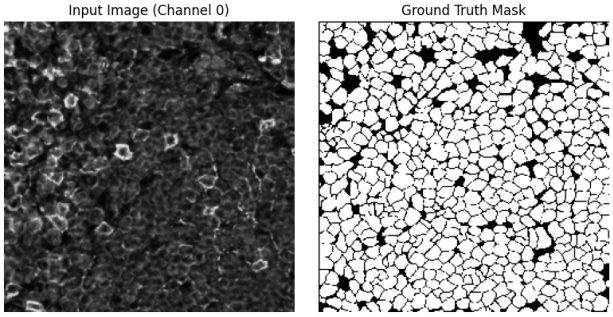

Figure 9. Sample input from Channel 0 (left) and corresponding ground truth segmentation mask (right)

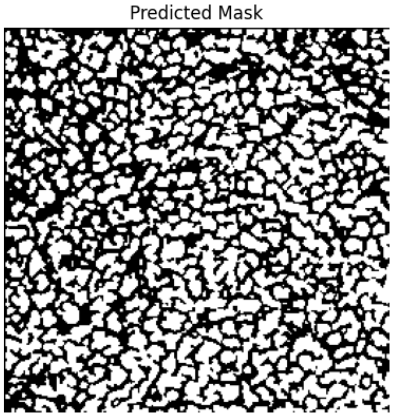

Figure 10. Model output showing successful nuclear detection but limited instance separation in high-density regions.

The visualization panel provides qualitative insight into performance. The input image shows a dense and noisy nuclear environment. The ground truth mask displays precise nuclei separation, including boundaries for tightly packed cells. The predicted mask, however, reveals challenges - although nuclei are detected globally, they appear merged or over-segmented. This aligns with the high F1-score (0.71) and precision (0.82), but lower Panoptic Quality (0.036) and $F_1^{0.5}$ score (0.034), indicating that instance separation is still a bottleneck.

Overall, the model demonstrates strong generalization across staining protocols and effective pixel-level classification but would benefit from enhanced post-processing or architectural improvements to address nuclear overlaps.

4. Ethical and Societal impact

While our model shows promising performance in nuclei segmentation from multiplexed biomedical images, several ethical and societal considerations remain important.

4.1. Dataset Diversity and Generalization

Medical imaging datasets, including multiplexed ones like CODEX, often lack demographic diversity and metadata such as ethnicity or age. This can lead to models that generalize poorly across populations, introducing performance disparities that may worsen healthcare inequality [17].

4.2. Risks in Clinical Deployment

Even accurate models can fail in real-world use due to differences in imaging conditions or rare edge cases. Mis-segmentation could affect diagnoses or downstream treatment planning. Without proper validation and monitoring, AI systems may be over-relied on, especially if outputs appear confident but are incorrect [18][19].

4.3. Transparency and Oversight

To ensure safe integration into clinical workflows, models must be developed and reported with transparency: clear documentation of training data, methods, and subgroup performance. Reproducibility and human-in-the-loop systems are essential to ensure clinicians can audit, override, or contextualize AI outputs as needed [20].

In summary, responsible AI in medical imaging must prioritize fairness, safety, and clinical collaboration to ensure equitable outcomes across all patient groups.

5. Project organization

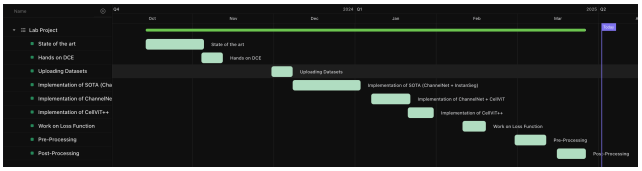

Figure 11. Gantt schema of our work organization

Over the past 6 months, we have been working on our Lab Project. It is a large-scale research project involving knowledge in medicine and Deep Learning.
To achieve this, we had to organize ourselves in such a way that we sometimes worked together on the same task, and sometimes divided the tasks to progress faster.
Initially, we both worked on the state of the art, the DCE, and the datasets.

Then, we both worked on the implementation of ChannelNet and InstanSeg. It was important that we both had solid foundations.

Maria then worked on the implementation of ChannelNet and CellViT, the loss functions, and post-processing. Zacharie worked on the implementation of CellViT++ and pre-processing.

6. Conclusion and Perspectives

6.1. Training Dynamics

Throughout the training process, the loss curves exhibited a steady convergence over 130 epochs, demonstrating the model's ability to learn features efficiently. The F1 Mean score plateaued around 0.35, indicating that the model achieved stable generalization without overfitting. This consistent performance highlights the effectiveness of our training strategy and the robustness of the architecture in handling diverse datasets.

6.2. Qualitative Results

Our model's predictions on challenging CODEX samples demonstrated strong shape and boundary detection, even in low-contrast areas. Despite slight over-segmentation in densely populated regions, the overall structure of the nuclei was well-preserved. These results underscore the model's capability to accurately segment nuclei in complex biomedical images.

The model achieved an F1 score of 0.71, a Precision of 0.82, and a mean Average Precision (mAP) of 0.77. These metrics reflect the model's high accuracy in detecting and segmenting nuclei, showcasing the potential of the hybrid CNN–Transformer approach for robust, stain-agnostic nuclei segmentation.

6.3. Future Work

While our model demonstrates solid performance in pixel-wise nuclear segmentation, several promising directions could be pursued to improve instance-level accuracy and overall generalization. Below, we outline key areas for further research and optimization:

Our current training set consists of only 65 samples, with 9 additional samples reserved for testing. Increasing the dataset size would enhance the model’s robustness, particularly in handling diverse nuclear morphologies and staining variations. This could be achieved by collecting more annotated data, leveraging data augmentation techniques, or employing synthetic data generation methods to enrich the training distribution.

Training deep learning models, especially transformer-based architectures, is computationally expensive and contributes to carbon emissions. Future efforts could explore strategies to minimize environmental impact, such as model pruning, knowledge distillation, or leveraging energy-efficient hardware (e.g., TPUs). Furthermore,

experimenting with smaller yet high-performing architectures, such as lightweight ViTs or efficient CNN backbones, could reduce computational costs while maintaining accuracy.

CellViT provides both a segmentation map and vector embeddings as outputs. These embeddings capture high-level semantic information about cell structures, which could be leveraged to refine segmentation accuracy. Potential approaches include:

- Using clustering methods (e.g., K-Means, DBSCAN) on embeddings to improve instance separation, particularly in crowded regions.
- Training a post-processing model (e.g., MLP, U-Net) that refines segmentation predictions based on both the segmentation map and the embeddings.
- Developing a verification network that classifies segmented regions using embeddings to eliminate false positives and correct segmentation errors.
- Integrating embeddings into a graph-based model (e.g., Graph Neural Networks - GNN) to better model relationships between nuclei and refine segmentation boundaries.

Beyond these improvements, post-processing could be enhanced by integrating the Segment Anything Model (SAM). Leveraging SAM's general-purpose segmentation capabilities could refine instance masks and improve boundary delineation, particularly in ambiguous or low-contrast regions.

Additionally, replacing the transformer encoder in CellViT with pretrained ViT weights from MedSAM may improve feature extraction, particularly in medical imaging contexts where domain-specific representations are advantageous. This adaptation could enhance learning efficiency and accuracy, especially in low-data regimes.

Finally, while this study focused on the CPDMI dataset, broader validation on publicly available benchmarks such as PanNuke or TissueNet would enable more comprehensive comparisons with state-of-the-art methods and help evaluate scalability across different tissue types and pathological conditions.

References

[1] Global Burden of Disease 2019 Cancer Collaboration (Kocarnik JM, et al.). Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups, 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. JAMA Oncol. 2022;8(3):420–444. https://doi.org/10.1001/jamaoncol.2021.6987

[2] [Wu J, Mayer AT, Li R, et al. Integrated imaging and molecular analysis to decipher tumor microenvironment in the era of immunotherapy. Semin Cancer Biol. 2022;84:310–328. https://doi.org/10.1016/j.semcancer.2020.12.005

[3] He L, Long LR, Antani S, Thoma GR. *Histology image analysis for carcinoma detection and grading*. *Comput Methods Programs Biomed.* 2012;107(3):538–556. https://doi.org/10.1016/j.cmpb.2011.12.007

[4] [Choi SK, Cho SI, Jung W, et al. Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer. NPJ Breast Cancer. 2023;9(1):71. https://doi.org/10.1038/s41523-023-00577-4

[5] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). *Clinical-grade computational pathology using weakly supervised deep learning on whole slide images*. *Nature Medicine*, 25(8), 1301–1309.

[6] Gamper, J., Koohbanani, N. A., Benes, K., Graham, S., Jahanifar, M., Khurram, S. A., Azam, A., Hewitt, K., & Rajpoot, N. (2020). *PanNuke dataset extension, insights and baselines*. *arXiv preprint* arXiv:2003.10778.

[7] Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). *Data-efficient and weakly supervised computational pathology on whole-slide images*. *Nature Biomedical Engineering*, 5(6), 555–570.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. *arXiv preprint* arXiv:2010.11929.

[9] Hörst, F., Ting, S., Liffers, S. T., Pomykala, K. L., Steiger, K., Albertsmeier, M., Angele, M. K., Lorenzen, S., Quante, M., Weichert, W., & Kleesiek, J. (2023). *Histology-based prediction of therapy response to neoadjuvant chemotherapy using deep learning*. *JCO Clinical Cancer Informatics*, 7, e2300038.

[10] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems, 34, 2136–2147.

[11] 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.

[12] Goldsborough, T., O'Callaghan, A., Inglis, F., Leplat, L., Filby, A., Bilen, H., & Bankhead, P. (2024). *A novel channel invariant architecture for the segmentation of cells and nuclei in multiplexed images using InstanSeg*. Preprint.

[13] Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., & Kleesiek, J. (2024). *CellViT: Vision Transformers for Precise Cell Segmentation and Classification*. *Medical Image Analysis*, 103143.

[14] Aleynick, N., Li, Y., Xie, Y., Zhang, M., Posner, A., Roshal, L., Pe'er, D., Vanguri, R. S., & Hollmann, T. J. (2023). *Cross-platform dataset of multiplex fluorescent cellular object image annotations*. *Scientific Data*, 10(1), 193. https://doi.org/10.1038/s41597-023-02108-z Aleynick, N., Li, Y., Xie, Y., Zhang, M., Posner, A., Roshal, L., Pe'er, D., Vanguri, R. S., & Hollmann, T. J. (2023). *Cross-platform dataset of multiplex fluorescent cellular object image annotations*. *Scientific Data*, 10(1), 193. https://doi.org/10.1038/s41597-023-02108-z

[15] Ziv, M., Gruber, G., Sharon, M., Vinogradov, E., & Yeger-Lotem, E. (2022). *The TissueNet v.3 Database: Protein-protein interactions in adult and embryonic human tissue contexts*. *Journal of Molecular Biology*, 434(11), 167532. https://doi.org/10.1016/j.jmb.2022.167532

[16] Berman, M., Rannen Triki, A., & Blaschko, M. B. (2018). The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4413–4421. https://doi.org/10.1109/CVPR.2018.00463

[17] Seyyed-Kalantari, L., Liu, G. Y., McDermott, M. B. A., Ghassemi, M., & Chen, I. Y. (2021). *Underdiagnosis bias of artificial intelligence algorithms in chest X-rays*. *Nature Medicine*, 27, 217–219. https://doi.org/10.1038/s41591-020-01209-7

[18] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Thomas, L. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. European Journal of Cancer, 113, 47–54. https://doi.org/10.1016/j.ejca.2019.04.001

[19] Oakden-Rayner, L., Beam, A. L., & Palmer, L. J. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of the National Academy of Sciences (PNAS), 117(48), 30033–30039. https://doi.org/10.1073/pnas.1917114117

[20] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310. https://doi.org/10.1186/s12911-020-01332-6