

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE  
EPFL-2025 - MACHINE LEARNING

## Nuclei Instance and Semantic Segmentation using Spatial Proteomics and H&E Data

*Students:*

MIZIA Michal  
KUCI Jon  
LI Huaqing

*Professor:*

WEST Robert

December 18, 2025



## 1 Introduction

Advances in pathology increasingly rely on deep learning models capable of extracting rich morphological and molecular information from tissue images. CellViT [5], a transformer-based encoder-decoder model for nuclei segmentation, provides a strong foundation for analyzing conventional H&E-stained light microscopy data, where the cell structure is visualized through chemical staining. In contrast, Spatial-Proteomics (SP) [8] imaging captures multiplexed protein expression at single-cell resolution, offering a complementary view of the same tissue. In this field, the Virtues encoder represents a state-of-the-art model for extracting semantically meaningful features from SP data [9].

This project explores the use and extension of the CellViT decoder architecture across these two domains using identically registered tissue sections, enabling a direct comparison between morphological (H&E) and molecular (SP) representations of the same underlying cellular architecture.

## 2 Dataset

The dataset [6] consists of 35 paired whole-slide images, each available in two imaging modalities: conventional H&E light microscopy and Spatial Proteomics (SP). Every WSI covers the same registered tissue area of size  $3000 \times 3000$  pixels, enabling direct pixel-level correspondence between morphological (H&E) and molecular (SP) signals.

For model development and visualization tasks, the first 28 WSIs (sorted alphabetically by tissue identifier) are used for training, while the remaining 7 WSIs are held out for validation on unseen tissue specimens.

### 2.1 Spatial Proteomics Images

The SP images of the given dataset contain 16 biomarker channels that measure protein expression levels. The SP panel includes the following markers: *CD31*, *CD45*, *CD68*, *CD4*, *FOXP3*, *CD8a*, *CD45RO*, *CD20*, *PD-L1*, *CD3e*, *CD163*, *E-cadherin*, *PD-1*, *Ki67*, *Pan-CK*, and *SMA*.

### 2.2 Nuclei Masks

Each WSI is accompanied by a dense, per-pixel nuclei-type annotation map with 10 classes. These classes are highly imbalanced, with relative frequencies as follows:

Table 1: Relative pixel frequency of nuclei classes in the dataset.

Background 76.7%	Stroma 8.4%	Tumor 10.3%	Macrophage 1.9%	Helper T cell 1.1%
Cytotoxic T cell 0.4%	Endothelia 0.6%	Other 0.5%	Treg 0.1%	B cells 0.007%

## 3 Methods

### 3.1 Model Overview

Our segmentation framework integrates a frozen VirTues encoder with an adapted CellViT-style decoder to produce dense nuclei-type predictions. An overview of the architecture is shown in Fig. 1. The encoder produces patch-scale semantic (PSS) tokens, which are progressively upsampled by the decoder to reconstruct full-resolution segmentation maps.

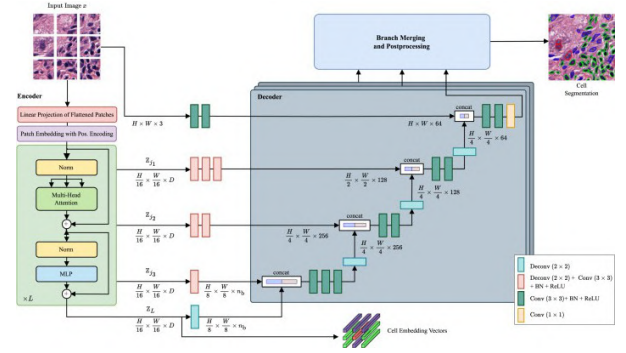


Figure 1: Overview of the proposed encoder-decoder architecture combining the VirTues encoder with a CellViT-derived decoder.

### 3.2 VirTues Encoder

For feature extraction, we use the VirTues encoder, consisting of 16 transformer blocks operating on  $8 \times 8$  image patches. The encoder is kept frozen during training. To enable compatibility with CellViT-style decoding, the encoder is modified to return intermediate hidden states at multiple depths, which act as skip connections.

### 3.3 Decoder

The decoder architecture is based on the CellViT design but adapted to process the higher-resolution 8-pixel patch embeddings produced by the VirTues encoder. Since CellViT assumes a 16-px patch size, directly inserting VirTues tokens would alter the expected spatial dimensionality. To address this, the decoder was reconfigured.

#### 3.3.1 Decoder Adaptation

To achieve an overall upsampling factor of  $8 \times$ , one transposed-convolution ( $2 \times 2$ ) layer was removed from each of the decoder's upsampling stages (indicated in red), as well as from the central bottleneck stage (cyan). Each removed transposed-convolution block was replaced with a  $1 \times 1$  convolutional layer, ensuring that spatial dimensionality was preserved while minimizing disruptions to the information flow and maintaining feature integrity.

The final decoder layer outputs dense nuclei-type segmentation maps corresponding to the 10 broad categories defined in the dataset.

## 4 Experiments

### 4.1 Data Splitting and Training Protocol

For the first three experiments, each Whole Slide Image (WSI) dataset was partitioned into smaller subsets to accelerate training:

- **Fold 1:** train = first 25% of all items, test = second 25%.
- **Fold 2:** train = third 25% of all items, test set = last 25%.

Training was run for 30 epochs and employed the AdamW optimizer [7] ( $5 \times 10^{-4}$  learning rate,  $1 \times 10^{-2}$  weight decay) with a cosine annealing warm restart scheduler ( $T_0 = 20$ ,  $\eta_{\min} = 10^{-6}$ ). Only in the augmentation experiment, we aimed for full 3-fold cross-validation, as we theorized that overly small train sets might not work well with data augmentation.

## 4.2 Loss Function Comparison

We evaluated the effect of different segmentation loss formulations on nuclei segmentation performance. The total segmentation loss used in our experiments combined CE (with optional class weights) and Dice losses, optionally with Focal Tversky loss. The Focal Tversky loss is a generalization of the Tversky loss that emphasizes underrepresented classes, improving performance on imbalanced datasets [1].

$$L_{\text{total}} = 0.4 \times L_{\text{CE}} + 0.4 \times L_{\text{Dice}} + \lambda_{\text{FT}} \times L_{\text{FT}} \quad (1)$$

where  $\lambda_{\text{FT}}$  was set to either 0.0 or 0.2 depending on the experiment.

Configuration	FT	CW	Average Dice
CE + Dice	0.0	No	0.4653 $\pm$ 0.0241
CE + Dice	0.0	Yes	0.4622 $\pm$ 0.0289
CE + Dice + FT	0.2	No	<b>0.4788 <math>\pm</math> 0.0431</b>
CE + Dice + FT	0.2	Yes	0.4644 $\pm$ 0.0294

Table 2: Loss function experiment results. FT: Focal Tversky weight; CW: class weights applied to CE loss. Dice scores reported as mean  $\pm$  std across 2 folds.

The best performing configuration was CE + Dice with FT weight 0.2, without class weighting, achieving an average Dice of 0.4788, corresponding to a +2.92% improvement over the baseline. Combining class weights with FT reduced performance compared to FT only, which suggests, that adding FT loss already focuses the model on underrepresented nuclei classes enough. This confirms that Focal Tversky loss helps in mitigating class imbalance.

## 4.3 Skip Connection

We evaluated the effect of incorporating skip connections from multiple encoder depths, as well as the original input image [Figure 1], compared to utilizing only the final encoder embeddings.

Configuration	Average Dice	Fold 1	Fold 2
SP-only, No Skip Conn	0.4330 $\pm$ 0.0295	0.4036	0.4625
SP-only, With Skip Conn	0.4787 $\pm$ 0.0492	0.4295	0.5280

Table 3: Average Dice scores for different skip connection usage.

The results indicate that the addition of skip connections substantially improves segmentation performance, increasing the average Dice score by more than 10% for SP-only data. Skip connections are essential for reconstruction of fine-grained spatial details from early layers to the decoder, which is particularly valuable for accurate segmentation of small objects, such as nuclei.

## 4.4 Oversampling Strategy

To address class imbalance in nuclei types, we evaluated a patch-level oversampling strategy adapted from CellViT[?]. For each training patch  $i$ , a sampling weight is computed based on the presence of underrepresented cell classes.

$$w_{\text{cell}}(i, \gamma_s) = (1 - \gamma_s) + \gamma_s \sum_{c=1}^C \frac{c_{i,c} N_{\text{cell}}^{\gamma_s}}{\gamma_s \sum_{k=1}^{N_{\text{train}}} c_{k,c} + (1 - \gamma_s) N_{\text{cell}}} \quad (2)$$

where  $c_{i,c} \in \{0, 1\}$  indicates whether class  $c$  is present in patch  $i$ ;  $N_{\text{cell}}$  is the total number of cell instances;  $N_{\text{train}}$  the number of training patches;  $C$  the number of nuclei classes; and  $\gamma_s \in [0, 1]$  controls the oversampling strength. Higher  $\gamma_s$  values give more weight to patches containing rare cell types.

Table 4: Summary of cell oversampling experiments with different  $\gamma_s$  values.

Experiment	Mean Dice	Std Dice
Oversampling $\gamma_s = 0.0$	0.4798	0.0624
Oversampling $\gamma_s = 0.45$	0.4743	0.0719
Oversampling $\gamma_s = 0.85$	0.4930	0.0778

These results demonstrate that moderate-to-strong oversampling improves the model’s ability to segment underrepresented nuclei classes, reducing fold-to-fold variability and accelerating convergence.

## 4.5 Augmentation Strategy

A challenge often observed during training was overfitting, characterized by a continued decrease in training loss while the validation loss plateaued. For this reason, we introduced a lightweight augmentation pipeline designed to improve generalization:

- **Geometric augmentations:** random horizontal flip, vertical flip, and 90° rotations.
- **Appearance augmentations (H&E only):** RGB Color jitter with probability=0.5.
- **CyCIF channel dropout:** randomly zeroing entire channels with probability=0.1 to improve robustness to signal variability.

Table 5: Augmentation experiment: Full 3-fold cross-validation results with 50 training epochs each

Experiment	Fold 1	Fold 2	Fold 3
No Augmentation	0.5811	0.5340	0.5322
With Augmentation	0.5947	0.5498	0.5367
<b>Mean <math>\pm</math> Std</b>	<b>0.5491 <math>\pm</math> 0.0226</b>	<b>0.5604 <math>\pm</math> 0.0249</b>	

Incorporating augmentations yielded a modest but consistent improvement in Dice score across folds, suggesting that light augmentations help regularize the model and improve generalization. Notably, the use of CyCIF channel dropout may increase robustness to variability in marker intensity across different acquisition conditions.

## 5 Encoder Comparison

### 5.1 Training Protocol

For sections 5 and 6, the encoders were kept frozen and only the decoder parameters were optimized. Precomputed embeddings

were used with a tissue-level sequential split (80% train, 20% validation), sorted alphabetically by tissue identifier.

Training employed the AdamW optimizer [7] ( $5 \times 10^{-4}$  learning rate,  $1 \times 10^{-2}$  weight decay) with a cosine annealing warm restart scheduler ( $T_0 = 20$ ,  $\eta_{\min} = 10^{-6}$ ). Models were trained for up to 100 epochs with early stopping patience of 30 and batch size 128 and the best validation checkpoint was used for evaluation. Each  $3000 \times 3000$  pixel SP image was split into 625 smaller images of size  $120 \times 120$  pixel to fit into GPU memory. A single, A100-SXM4-80GB Nvidia GPU was used for training.

## 5.2 Semantic Segmentation

In this experiment, we directly compared the representational quality of the **CellViT encoder** [5] and the **Virtues encoder** [9] for semantic nuclei segmentation. To ensure a fair comparison, the *exact same decoder architecture* was used in both cases, with only a minimal modification to accommodate differences in patch size. Both models were trained using the same loss configuration (CE + Dice + Focal Tversky with weight 0.2), without oversampling or data augmentation.

The Virtues-based model substantially outperformed the CellViT-based model, achieving a best Dice score of 0.6293 compared to 0.2696 for the CellViT encoder. This large performance gap is further explained by the class-wise Dice scores shown in Table 6.

Table 6: Class-wise predicted percentage (Pred), ground-truth percentage (GT), and Dice score for CellViT and Virtues encoders using the same decoder.

Class	GT (%)	CellViT		Virtues	
		Pred (%)	Dice	Pred (%)	Dice
0	67.47	67.67	0.9092	64.81	0.8888
1	10.40	10.90	0.5102	11.29	0.5530
2	0.97	0.00	0.0000	1.21	0.5754
3	8.72	10.13	0.4966	7.88	0.5833
4	7.12	8.07	0.7325	8.69	0.7905
5	2.30	3.23	0.2211	2.64	0.5963
6	1.46	0.00	0.0000	1.41	0.6563
7	0.83	0.00	0.0000	0.87	0.3457
8	0.37	0.00	0.0000	0.72	0.5117
9	0.37	0.00	0.0000	0.48	0.5415

While both encoders performed well on the dominant background class (Class 0), the CellViT encoder failed to detect several rare nuclei classes entirely. In contrast, the Virtues encoder produced non-zero Dice scores across all classes and exhibited a markedly better alignment between predicted and ground truth class distributions.

These results indicate that SP data acts as an implicit form of regularization for semantic segmentation, encouraging the learning of more semantically meaningful and class-discriminative representations. Notably, this advantage persists even though both encoders were trained using the same self-supervised DINO [2] objective, highlighting the importance of the input modality itself rather than the training strategy alone.

## 5.3 Instance Segmentation

Instance segmentation requires separating individual nuclei instances, including touching or overlapping cells, and is a more difficult task than semantic segmentation. CellViT follows the HoVer-Net paradigm and demonstrates that high-quality instance segmentation can be achieved using a watershed algo-

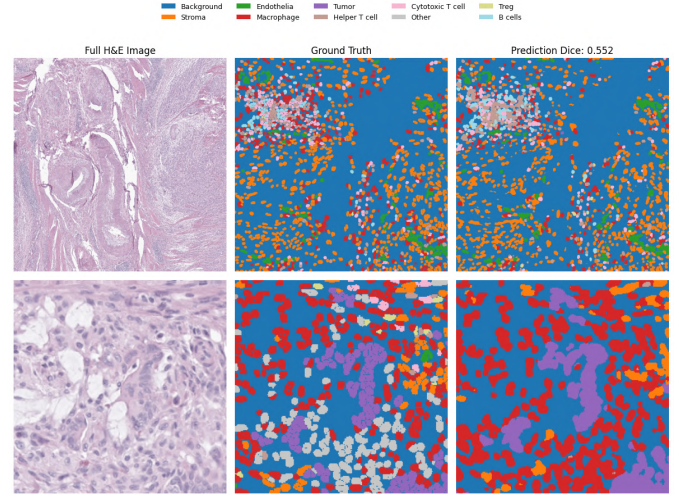


Figure 2: Predicted and ground-truth semantic segmentation maps for Virtues (top) and CellViT (bottom) encoders. Note that because of differences in patch size used by the encoders, the final image resolutions are  $256 \times 256$  for CellViT and  $600 \times 600$  for VirTues. We can see that CellViT maps do not predict any rare cell types.

rithm, on two complementary pixel-wise representations: a binary nuclei map and horizontal-vertical (HV) distance map, which encodes signed horizontal and vertical offsets from each nucleus pixel to its corresponding nucleus center of mass, while pixels outside nuclei are ignored.

### 5.3.1 Training Objective

To evaluate the ability of different encoders to produce informative binary and HV representations, we adopt the original CellViT loss formulation restricted to these two branches. The total training loss is defined as

$$L_{\text{total}} = \lambda_{\text{FT}} L_{\text{FT}} + \lambda_{\text{Dice}} L_{\text{Dice}} + \lambda_{\text{MSE}} L_{\text{MSE}} + \lambda_{\text{MSGE}} L_{\text{MSGE}}, \quad (3)$$

where Focal Tversky and Dice losses supervise the binary nuclei prediction, while mean squared error (MSE) and mean squared gradient error (MSGE) enforce accurate and spatially coherent HV maps.

### 5.3.2 Encoder Comparison on HV and Binary Maps

The training was done according to Section 5.1 with the same loss for both decoders.

Table 7 summarizes the resulting performance on the HV and binary branches.

Table 7: Performance of CellViT and Virtues encoders for HV and binary map prediction. Average HV MSE, Dice score on the binary map, and best total loss on validation set are reported.

Encoder	Avg HV MSE	Avg Dice	Best Loss
CellViT (H&E)	0.0566	0.7170	0.4178
Virtues (SP)	0.0743	0.7156	0.4827

Interestingly, the results show that H&E inputs slightly outperform SP data for the HV and binary map branches, though the difference is marginal. This is plausible because



H&E images provide strong structural cues such as clear nucleus boundaries, which are highly informative for predicting pixel-wise HV distances and binary segmentation. In contrast, SP data emphasizes protein expression patterns that are more relevant for semantic or class-discriminative features. Overall, this demonstrates that both modalities can produce high-quality instance segmentation maps.

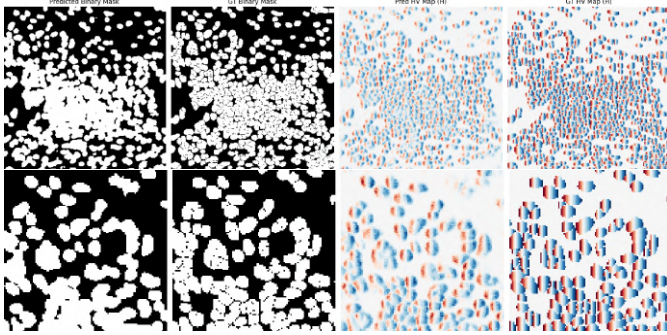


Figure 3: Predicted and ground-truth binary and HV maps for CellViT (top) and Virtues (bottom) encoders. Note that because of differences in patch size used by the encoders, the final image resolutions are 256x256 for CellViT and 120x120 for VirTues.

## 6 CellViT Architectural Variants and Extensions

We investigate decoder-level architectural extensions to the CellViT framework when operating on Spatial Proteomics (SP) embeddings produced by a frozen VirTues encoder. All experiments use the same encoder, data split, and training protocol, and differ only in the decoder architecture, allowing a controlled evaluation of decoder-level inductive biases.

### 6.1 Boundary Supervision

We extended the baseline decoder with an auxiliary boundary prediction head attached to the final decoder stage. The head predicts a single-channel boundary map and is trained jointly with the nuclei-type segmentation head while sharing the same decoder backbone.

Boundary supervision provides an explicit geometric signal that improves separation of touching nuclei while introducing negligible computational overhead. This configuration is therefore used as the reference decoder for subsequent extensions.

### 6.2 Global Context Modeling

As an alternative to mask-restricted attention, we add a global context block at the decoder bottleneck that applies self-attention over spatial tokens after bottleneck upsampling.

This modification introduces global spatial interactions at the bottleneck but increases memory consumption due to the quadratic complexity of global attention.

### 6.3 Masked Self-Attention

Motivated by mask-restricted attention mechanisms proposed in Masked2Former [3], we introduce a masked self-attention module at the final decoder stage. Attention is restricted to foreground regions using a predicted foreground mask to reduce background feature mixing during late-stage decoding.

Table 8: Performance of CellViT decoder architectural variants on SP-only embeddings. Best Dice and mean Dice  $\pm$  standard deviation are reported over training epochs.

Configuration	Best Dice	Mean Dice $\pm$ Std
SP only (baseline)	0.6255	0.6065 $\pm$ 0.0299
SP + boundary supervision	0.6291	<b>0.6106 <math>\pm</math> 0.0312</b>
SP + boundary + global context	0.6275	0.6075 $\pm$ 0.0306
SP + boundary + masked attention	<b>0.6309</b>	0.6104 $\pm$ 0.0299

Masked self-attention introduces additional computational and memory overhead due to attention computation on dense feature maps.

## 6.4 Training Protocol

The training configuration for the decoder architectural variants is identical to the protocol described in Section 5.1. Due to time and computational constraints, each configuration was evaluated using a single fold.

## 6.5 Results

Table 8 summarizes the performance of all decoder variants on SP-only embeddings. Boundary supervision provides consistent improvements over the baseline with low variance, making it the most robust modification. Masked self-attention achieves the highest peak Dice score but at increased computational and memory cost. Global context modeling yields stable predictions but does not improve performance relative to boundary supervision.

## 7 Conclusion

In this work, we analyzed design choices for nuclei segmentation across semantic and instance-level tasks. We showed that architectural components such as multi-scale skip connections substantially improve segmentation quality by preserving fine-grained spatial information, while the inclusion of Focal Tversky loss effectively mitigates class imbalance without requiring class weighting. A key finding is the strong impact of input modality on semantic segmentation performance: the Virtues encoder trained on spatial proteomics data significantly outperformed the CellViT encoder on rare and minority cell types, despite both being trained with the same self-supervised DINO objective. In contrast, instance segmentation performance, evaluated through binary and HV distance maps, was comparable across modalities. These results suggest that spatial proteomics data provides more semantically meaningful representations, while H&E information remains sufficient for geometry-driven analysis. Future work may explore hybrid models combining SP and H&E modalities, end-to-end encoder finetuning, and larger dataset sizes. At the decoder level, lightweight architectural extensions such as auxiliary boundary supervision provided consistent gains with minimal overhead, while attention-based mechanisms offered higher peak performance at increased computational cost. Overall, this study highlights the importance of modality choice in representation learning for computational pathology and demonstrates the complementary strengths of molecular and histological imaging.

## 8 Ethical Risk Assessment

### 8.1 Risk

Errors in semantic segmentation of SP and H&E images could lead to misidentification of cell types or tissue regions. In a clinical setting, this may influence biomarker discovery or treatment decisions, potentially affecting the well-being of patients.

Misclassification of rare but clinically relevant cell types could result in incorrect conclusions about tumor environments and recommended treatments. The severity is high because decisions based on flawed data could misinform clinical strategies. Likelihood is moderate to high, as segmentation models generally perform well on abundant cell types but struggle with rare populations.

### 8.2 Evaluation

We quantified model performance using class-wise Dice scores and compared predicted versus ground truth distributions. Common failure modes were rare cell types and low-contrast regions.

### 8.3 Mitigation

We emphasize that models are intended for research purposes only and are not validated for direct clinical decision-making. Performance limitations for rare cell types are documented. Additionally, in the case of unadvised clinical usage, we recommend contrasting the model predictions with an independent human expert to not overrely on one source of information.

## References

- [1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. *arXiv preprint arXiv:1810.07842*, 2019.
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.
- [3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022.
- [4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.
- [5] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024.
- [6] JR Lin, YA Chen, D Campton, and et al. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature Cancer*, 4:1036–1052, 2023.
- [7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
- [8] Andreas Mund, Andreas-David Brunner, and Matthias Mann. Unbiased spatial proteomics with single-cell resolution in tissues. *Molecular Cell*, 82(12):2335–2349, 2022.
- [9] Johann Wenckstern, Eeshaan Jain, Yexiang Cheng, Benedikt von Quersurth, Kiril Vasilev, Matteo Pariset, Phil F. Cheng, Petros Liakopoulos, Olivier Michielin, Andreas Wicki, Gabriele Gut, and Charlotte Bunne. AI-powered virtual tissues from spatial proteomics for clinical diagnostics and biomedical discovery. *arXiv preprint arXiv:2501.06039*, 2025.