Deep Learning-Based Nuclei Segmentation for Label-Free Histology Using Photon Absorption Remote Sensing Microscopy

Gloria J. So[1] Umar Ali[1] James A. Tummon Simmons[1]
James E. D. Tweel[1] Benjamin R. Ecclestone[1] Parsin Haji Reza[1]

[1]PhotoMedicine Labs, Systems Design Engineering, University of Waterloo, Ontario, N2L 3G1, Canada

gloria.so@uwaterloo.ca, umar.ali@uwaterloo.ca, jatummonsimmons@uwaterloo.ca,
james.tweel@uwaterloo.ca, benjamin.ecclestone@uwaterloo.ca, parsin.hajireza@uwaterloo.ca

Abstract

As pathology workflows shift toward digitization, deep learning is increasingly being used for tasks such as nuclei segmentation. Photon Absorption Remote Sensing (PARS) is a label-free imaging technique that provides nuclear and extranuclear contrast. PARS images can be used to generate virtual H&E (VHE) images that emulate conventional H&E staining. Although many nuclei segmentation algorithms exist for conventional histology, none have yet been applied to PARS. We evaluated three state-of-the-art deep learning models for nuclei segmentation: StarDist, Cellpose, and DeepCMorph. We demonstrated that pretrained models show consistent interpretation across H&E and VHE, with a mean percent error of 1.54% in nuclear count. We also trained models directly on PARS for the first time. The best model produced masks that preserved cell morphology, with errors of only -3.96% in nuclear count, -0.29% in intercellular distance, and 1.53% in nuclear area. These results show PARS as an effective technique for automated nuclear analysis in label-free histopathology.

1. Introduction

Nuclei segmentation is a key task in digital pathology, allowing the extraction of features such as nuclear size, shape, and density from histological images. These features support tumour grading and identification of malignant characteristics in tissue [1]. Hematoxylin and eosin (H&E) staining is the gold standard for contrast, with hematoxylin staining nuclei dark blue-purple and eosin staining the cytoplasm and extracellular matrix pink [2–4]. However, staining is resource-intensive and prone to variability, motivating interest in label-free imaging methods that preserve tissue integrity [5].

Photon Absorption Remote Sensing (PARS) is a label-free modality which employs ultraviolet excitation to simultaneously capture the non-radiative thermal and radiative autofluorescence response of tissue [6]. In PARS imaging, nuclear contrast is provided primarily from the non-radiative relaxation of DNA, comparable to the contrast achieved through hematoxylin staining [7]. Conversely, radiative contrast highlights extranuclear structures, similar to the contrast provided by eosin staining [8]. Tweel et al. developed a virtual staining workflow to generate H&E-like images from PARS images [7]. Clinical concordance has been shown between PARS virtual H&E (VHE) and gold standard H&E in breast needle biopsies; this was further supported with nuclear statistics [6, 9].

Several deep learning nuclei segmentation models have been proposed, including StarDist [10–12], Cellpose [13, 14], and DeepCMorph [15]. All three use the U-Net architecture, but differ in how they generate nuclei masks. StarDist predicts star-convex polygons for each nucleus and resolves overlaps with non-maximal suppression [10]. Cellpose outputs a vector flow field that is converted into instance masks using a watershed approach [13]. DeepCMorph uses a segmentation module with an EfficientNet-B7 encoder that produces semantic probability maps. It also has a separate cell type classification module [15].

In this study, we selected three models widely used for nuclei segmentation in traditional histology: StarDist [10], Cellpose [13, 14], and DeepCMorph [15]. We trained and evaluated the models on H&E, VHE, and PARS to compare segmentation performance across the modalities. Pretrained models achieved comparable performance on H&E and VHE images, with a mean percent error of 1.54% in nuclear count, suggesting that deep learning models interpret the two modalities similarly. Additionally, the first nuclei segmentation models for PARS images were developed. The best model achieved mean per-tile errors of only -3.96% in nuclear count, -0.29% in intercellular distance, and 1.53% in nuclear area.

2. Methods

2.1. Data Acquisition

Two whole slide images of kidney tissue were used in this study. For each slide, three representations from the same tissue region were acquired: an H&E-stained image, a label-free PARS microscopy image, and a VHE image generated from PARS data. PARS images were captured following the procedure outlined in [8], and VHE images were produced according to [16].

2.2. Data Preprocessing

Precisely paired images are required for training supervised segmentation models. Weakly registered H&E, PARS, and VHE whole-slide images were divided into 2,540 tiles of 536x536 pixels, with a 12-pixel overlap on all sides. The overlap ensured proper registration near tile edges and was removed afterward with a 512x512 centre crop. Each H&E tile underwent two separate registrations: one to its corresponding VHE tile, producing a VHE-registered reference, and another to its corresponding PARS tile, producing a PARS-registered reference. Registering the H&E tiles to each modality preserves the native structure of the VHE and PARS data, allowing trained models to perform more accurately on real, unaltered images.

For fine registration, three-channel (RGB) H&E and VHE tiles were converted to grayscale by averaging across RGB channels. PARS tiles, consisting of the non-radiative and radiative channels, were converted to two-dimensional grayscale tiles using a weighted average of 0.2 for the radiative channel and 0.8 for the non-radiative channel. This weighting emphasizes the higher nuclear contrast in the non-radiative channel for fine registration, while enabling the algorithm to register coarser features from the radiative channel.

The registration used the Mattes Mutual Information metric with regular step gradient descent. Multi-resolution registration was applied across two levels. First, at the coarse level, each tile was downsampled by a factor of two and smoothed, and a linear interpolator was used to apply an affine transformation for initial alignment. Then, at the fine level, the RGB channels of each H&E tile were individually resampled and warped to their corresponding targets using a B-spline transformation. The channels were recombined to produce a registered H&E tile, which was cropped to 512x512 to remove border artifacts.

2.3. Experiments

Exp 1: Pretrained Model Evaluation on VHE. Pretrained nuclei segmentation models, StarDist's 2D Versatile H&E [10], Cellpose's cyto3 [14], and DeepCMorph's Pan_Cancer_32_classes_acc_827 segmentation module [15] were applied directly to VHE tiles and their registered H&E counterparts without fine-tuning or retraining. For each model, performance on VHE images was compared directly to the model's performance on registered H&E images. This assesses each model's robustness to differences between H&E and VHE images.

Exp 2: Model Training on VHE and PARS. The models were trained from scratch on VHE and PARS tiles using identical 80%-10%-10% splits for training, validation, and testing. The tiles were treated as independent samples.

Pseudo-ground truth masks were generated from fine-registered H&E tiles with StarDist's 2D Versatile H&E algorithm using a threshold of 0.3. This model was selected for its training on large and diverse H&E datasets [12], its use in prior VHE and PARS studies [6], and because validation on a subset of H&E tiles showed it outperformed Cellpose and DeepCMorph.

After training, model thresholds were optimized on a validation set. The L-BFGS-B algorithm [17] was used to optimize the Dice score (Eq. (1)). Dice was an evaluation metric used in the DeepCMorph paper [15] and is suitable for evaluating semantic segmentation outputs. The search bounds were: StarDist ($p \in [0.1, 0.9]$, $n \in [0.3, 0.5]$), Cellpose ($p \in [-2.5, 3]$, $f \in [0, 2]$), and DeepCMorph ($p \in [0.05, 0.9]$), where p is the probability threshold for all models, n the non-maximum suppression threshold for StarDist, and f the flow threshold for Cellpose.

$$\text{Dice} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{1}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

2.4. Evaluation Metric

A test set of $n = 254$ tiles, which were not used during training, was employed for evaluation. Segmentation performance was quantified using average precision (AP_τ, Eq. (3)) at different intersection over union (IoU, Eq. (2)) thresholds. This is a common metric for evaluating instance segmentation [10, 13]. IoU measures the overlap between a predicted nucleus mask, I_{pred}, and the ground truth nucleus mask, I_{gt}. For a given threshold τ, a predicted nucleus is considered a true positive (TP_τ) if its IoU with a ground truth nucleus exceeds τ. Unmatched predicted nuclei are false positives (FP_τ), and unmatched ground truth nuclei are false negatives (FN_τ).

$$\text{IoU} = \frac{I_{\text{pred}} \cap I_{\text{gt}}}{I_{\text{pred}} \cup I_{\text{gt}}} \tag{2}$$

$$\text{AP}_\tau = \frac{\text{TP}_\tau}{\text{TP}_\tau + \text{FP}_\tau + \text{FN}_\tau} \tag{3}$$

Additionally, cell-based metrics, including nuclear count, mean intercellular distance (ICD), and mean nuclear

area, were extracted from the predicted segmentation masks and compared to ground truth masks.

3. Results

Validating Fine-Registration. Differences between the fine-registered and original H&E images were assessed to ensure the fine-registration did not significantly distort key metrics. StarDist's default pretrained model [10] was used to segment nuclei from the original H&E tiles, the VHE-registered H&E tiles, and the PARS-registered H&E tiles. Table 1 displays paired confidence intervals (CIs) comparing nuclear count, mean intercellular distance (ICD), and mean nuclear area across regions.

Table 1. Per-tile mean differences and 95% CIs for cell metrics between original, VHE-registered, and PARS-registered H&E.

Metric	VHE-Registered	PARS-Registered
Nuclear Count	0.33 (-0.03, 0.69)	0.26 (-0.11, 0.62)
Mean ICD (μm)	-0.03 (-0.10, 0.04)	-0.03 (-0.10, 0.04)
Mean Area (μm^2)	-0.08 (-0.12, -0.05)	-0.09 (-0.13, -0.06)

For both VHE-registered and PARS-registered H&E images, differences in mean cell count and ICD relative to the original H&E were minimal. The 95% confidence intervals for the paired differences include zero, indicating insufficient evidence of a statistically significant difference. Mean nuclear area showed a small but statistically significant negative bias (< 0.1 μm^2); given that area scales with the square of radius, this bias is negligible compared to the typical nuclear size. These results suggest that the fine registration preserved image fidelity, supporting the use of the VHE-registered and PARS-registered H&E images as ground truth references in subsequent analyses.

Exp 1: Pretrained models on VHE. Table 2 summarizes per-tile percent errors and 95% confidence intervals for nuclear count, mean ICD, and mean nuclear area, comparing each pretrained model's predictions on VHE versus registered H&E. Absolute nuclear counts differed between models, but each model showed consistency between its prediction on H&E and its prediction on VHE. In terms of nuclear count, Cellpose showed no significant difference between H&E and VHE, but StarDist and DeepCMorph did. Despite this, the percent errors were small, at 1.54%, -0.30%, and -3.44% for StarDist, Cellpose, and DeepCMorph, respectively. For ICD, none of the models showed a statistically significant difference on VHE compared to H&E. Mean nuclear area showed statistically significant differences for all models, but percent errors remained small.

Table 2. Per-tile percent errors and 95% CIs for cell metrics from each pretrained model on VHE versus registered H&E.

Metric	StarDist	Cellpose	DeepCMorph
Nuclear Count	1.54 (0.42, 2.66)	-0.30 (-1.48, 0.88)	-3.44 (-4.42, -2.46)
Mean ICD	0.28 (-0.04, 0.60)	0.06 (-0.23, 0.34)	-0.18 (-0.56, 0.20)
Mean Area	-0.96 (-1.49, -0.43)	-2.18 (-2.83, -1.52)	1.33 (0.16, 2.50)

Exp 2: VHE- and PARS-trained models. Fig. 1 presents an example of a segmentation output from the models trained on VHE and PARS images. In the PARS images, even when nuclei were correctly identified, the boundaries between the ground truth and prediction showed wider discrepancies compared to VHE images.

The pseudo-ground truth masks occasionally contain errors, such as omitting faint or overlapping nuclei or incorrectly labelling non-nuclear regions. Using StarDist-generated masks also introduces a bias toward StarDist-style segmentations. Future work could explore probabilistic or consensus-based ground truth that integrates predictions from multiple models, rather than relying on a single model's binary prediction. Extensions of this work could also explore weak supervision, sparse annotations, or self-training methods that do not require precise pixel-level labels.

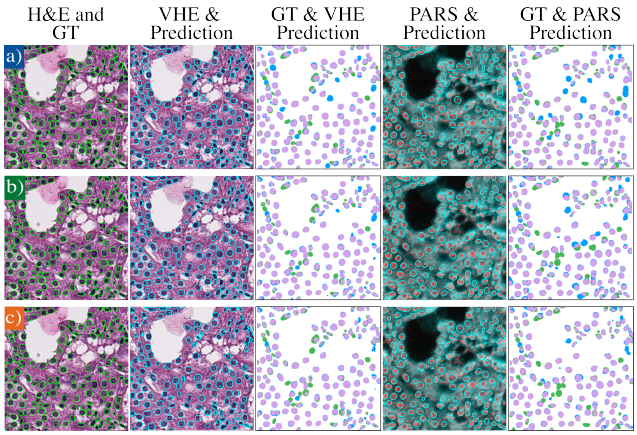

Figure 1. Segmentation masks for VHE- and PARS-trained models: a) StarDist, b) Cellpose, and c) DeepCMorph. Columns 3 and 5 show errors and agreement (green = false negative, blue = false positive, purple = true positive; more purple indicates better agreement between predictions and ground truth)

For the VHE-trained models, relative to the pseudo-ground truth, StarDist underestimated nuclear count by 2.92%, while Cellpose and DeepCMorph slightly overestimated by 2.53% and 0.67%, respectively. PARS-trained models all underestimated nuclear count, with DeepC-

Morph showing the largest bias (-9.83%), followed by Cellpose (-8.78%) and StarDist (-3.96%). For ICD, StarDist showed no significant difference compared to the ground truth on VHE, whereas the differences for Cellpose and DeepCMorph were statistically significant. On PARS, only Cellpose showed no significant difference. Mean area bias varied by model and imaging modality. Since nuclear area scales with the square of the radius, even small segmentation errors can cause amplified differences in mean area. VHE-trained models each showed only small differences, all within $\pm 1\ \mu\mathrm{m}^2$. PARS-trained models generally overestimated nuclear area compared to VHE-trained models, with the exception of StarDist, which achieved the smallest bias across all models. Challenges in nuclei detection and area estimation were more pronounced in PARS-trained models, especially with DeepCMorph and Cellpose.

Table 3. Per-tile mean % errors and 95% CIs for cell metrics from VHE-trained models versus pseudo-ground truth.

Metric	StarDist	Cellpose	DeepCMorph
Nuclear Count	-2.92 (-3.62, -2.22)	2.53 (1.76, 3.30)	0.67 (-1.12, 2.45)
Mean ICD	0.15 (-0.07, 0.38)	0.60 (0.36, 0.83)	3.36 (2.85, 3.88)
Mean Area	4.07 (3.48, 4.67)	-1.87 (-2.53, -1.22)	2.52 (1.00, 4.04)

Table 4. Per-tile mean % errors and 95% CIs for cell metrics from PARS-trained models versus pseudo-ground truth.

Metric	StarDist	Cellpose	DeepCMorph
Nuclear Count	-3.96 (-4.82, -3.10)	-8.78 (-9.63, -7.94)	-9.83 (-11.68, -7.97)
Mean ICD	-0.29 (-0.53, -0.04)	0.10 (-0.17, 0.37)	3.61 (3.02, 4.20)
Mean Area	1.53 (0.93, 2.14)	9.08 (8.36, 9.79)	15.25 (13.56, 16.94)

Fig. 2 shows the average precision (AP) for the pretrained, VHE-trained, and PARS-trained models. Our VHE- and PARS-trained models were optimized by Dice score rather than AP. Hence, these models are expected to underperform in the current evaluation when compared to literature-reported values for StarDist and Cellpose models, which were optimized by AP across IoU thresholds.

VHE-trained models achieved segmentation performance similar to that of pretrained models applied to VHE. Cellpose showed the largest improvement from its pretrained performance after VHE training, learning to better match the StarDist-generated pseudo-ground truth. After training, AP curves for StarDist and Cellpose are very close, indicating that any bias introduced by using StarDist masks is small and can be reduced with training. In general, tissue- and domain-specific training may have also improved the models' ability to adapt to the structural and contrast characteristics of VHE.

The PARS-trained models had slightly lower curves

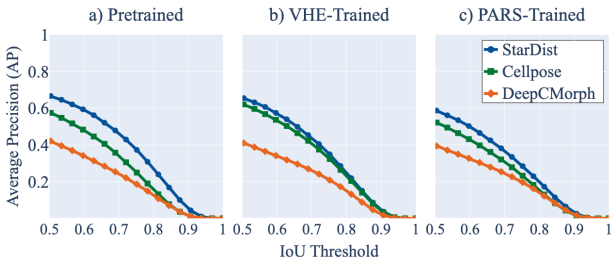

Figure 2. Average Precision for a) Pretrained models applied to VHE, b) VHE-Trained models applied to VHE, c) PARS-trained models applied to PARS. All predictions are compared to pseudo-ground truth masks.

compared to the VHE-trained models. These discrepancies could partly stem from registration inaccuracies during preprocessing, as initial misalignment appeared to be more pronounced in PARS images than in VHE images. DeepCMorph consistently showed the lowest curve since its semantic output required connected-component labeling to produce instance segmentations. Ultimately, the best PARS-trained model, StarDist, achieved an average precision comparable to its pretrained and VHE-trained counterpart. Its curve in Figure 2 looks similar to the results reported for StarDist on fluorescent cell images in [13]. Expanding the training dataset and optimizing thresholds based on Average Precision rather than Dice could further enhance segmentation performance.

Since VHE is derived from PARS via a CycleGAN, we know that meaningful nuclear features exist in PARS images [16]. Future work could explore alternative architectures, such as transformers, or leverage feature representations learned by the virtual staining model to improve PARS segmentation.

4. Conclusion

This study compares StarDist, Cellpose, and DeepCMorph on VHE and PARS images using average precision and cell-level metrics. Experiment 1 suggests that pretrained models perform similarly on VHE and H&E images, confirming VHE as a reliable, label-free substitute for conventional H&E. Experiment 2 showed that while VHE-trained models performed slightly better, PARS-trained models achieved competitive results. This demonstrates that direct segmentation on PARS is a viable strategy, especially when data is insufficient for training virtual staining models.

Overall, virtual staining enables accurate, label-free nuclear analysis with existing tools, while direct PARS provides comparable nuclear information with less computational overhead for training. These results demonstrate the potential of label-free PARS imaging for accurate nuclear analysis in digital pathology.

References

[1] Paari Murugan, Robert W. Allan, Lara R. Harik, Gladell P. Paner, John R. Srigley, Mahul B. Amin, Sean R. Williamson, Ying-Bei Chen, Kiril Trpkov, Michelle S. Hirsch, Steven C. Smith, Holger Moch, Peter A. Humphrey, Maria Carlo, and Viraj A. Master. Protocol for the Examination of Resection Specimens from Patients with Renal Cell Carcinoma, June 2024. 1

[2] Global Transformational Health Research Team, Frost & Sullivan. Global tissue diagnostics market, forecast to 2022. Market research report, Frost & Sullivan, 2018. Frost & Sullivan 2018. 1

[3] M. Slaoui and L. Fiette. Histopathology procedures: From tissue sampling to histopathological evaluation. In J.-C. Gautier, editor, *Drug Safety Evaluation: Methods and Protocols*, pages 69–82. Humana Press, Totowa, NJ, 2011.

[4] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller. Hematoxylin and eosin staining of tissue and cell sections. *CSH Protoc*, 2008:pdb.prot4986, May 2008. 1

[5] K. de Haan, Y. Zhang, J.E. Zuckerman, and et al. Deep learning-based transformation of h&e stained tissues into special stains. *Nature Communications*, 12:4884, 2021. 1

[6] James E. D. Tweel, Benjamin R. Ecclestone, Marian Boktor, Deepak Dinakaran, John R. Mackey, and Parsin Haji Reza. Automated Whole Slide Imaging for Label-Free Histology Using Photon Absorption Remote Sensing Microscopy. *IEEE transactions on bio-medical engineering*, 71(6):1901–1912, June 2024. 1, 2

[7] Benjamin R. Ecclestone, James E. D. Tweel, Deepak Dinakaran, John Mackey, and Parsin Haji Reza. Virtual histology with photon absorption remote sensing (PARS). In *Label-Free Biomedical Imaging and Sensing (LBIS) 2024*, volume PC12854, page PC128540L. SPIE, March 2024. 1

[8] Benjamin R. Ecclestone, Kevan Bell, Sarah Sparkes, Deepak Dinakaran, John R. Mackey, and Parsin Haji Reza. Label-free complete absorption microscopy using second generation photoacoustic remote sensing. *Scientific Reports*, 12(1):8464, May 2022. 1, 2

[9] James E. D. Tweel, Benjamin R. Ecclestone, Hager Gaouda, Deepak Dinakaran, Michael P. Wallace, Gilbert Bigras, John R. Mackey, and Parsin Haji Reza. Photon Absorption Remote Sensing Imaging of Breast Needle Core Biopsies Is Diagnostically Equivalent to Gold Standard H&E Histologic Assessment. *Current Oncology*, 30(11):9760–9771, November 2023. 1

[10] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell Detection with Star-Convex Polygons. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 265–273, Cham, 2018. Springer International Publishing. 1, 2, 3

[11] Martin Weigert and Uwe Schmidt. Nuclei Instance Segmentation and Classification in Histopathology Images with Stardist. In *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, pages 1–4, March 2022.

[12] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3655–3662, Snowmass Village, CO, USA, March 2020. IEEE. 1, 2

[13] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, January 2021. 1, 2, 4

[14] Carsen Stringer and Marius Pachitariu. Cellpose3: One-click image restoration for improved cellular segmentation, February 2024. 1, 2

[15] Andrey Ignatov, Josephine Yates, and Valentina Boeva. Histopathological Image Classification with Cell Morphology Aware Deep Neural Networks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6913–6925. IEEE Computer Society, June 2024. 1, 2

[16] James E. D. Tweel, Benjamin R. Ecclestone, Marian Boktor, James Alexander Tummon Simmons, Paul Fieguth, and Parsin Haji Reza. Virtual Histology with Photon Absorption Remote Sensing using a Cycle-Consistent Generative Adversarial Network with Weakly Registered Pairs, June 2023. 2, 4

[17] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, September 1995. 2